UNIVERSIDAD DE GRANADA

Facultad de Ciencias

Doble Grado en Física y Matemáticas

Trabajo Fin de Grado

Técnicas de Aprendizaje Supervisado Aplicadas a Clasificación de Topologías de Estado Final en el Experimento SBND

Presentado por:
D./D[a]. Luis Hallamaa Carmona

Curso Académico 2022-2023

Tutores:
D. José Luis Romero Béjar
D. Bruno Zamorano García

Declaración de originalidad

D. Luis Hallamaa Carmona

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2022-2023, es original, entendida esta, en el sentido de que no ha utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 11 de julio de 2023

Fdo: Luis Hallamaa Carmona

Summary/Resumen

Neutrinos are elementary particles with very small mass that interact weakly. Their experimental confirmation came in the mid-20th century, and since then, numerous dedicated experiments have been carried out to determine their characteristics and discover the mechanisms that govern their nature at a fundamental level. In this regard, the Short-Baseline Neutrino (SBN) program, an ambitious project whose name derives from the acronym for Short-Baseline Neutrino program, aims to resolve certain inconsistencies between measurements from previous experiments and thus tip the balance towards the existence or nonexistence of a new type of neutrino.

The importance of classifying final state topologies, that is, the particles observed after scattering with the nucleus and what can be inferred from them, lies in the utility of having categorized events for statistical analysis purposes. Different topologies are better explained by specific physical processes, which can be known with varying precision at both the theoretical and experimental levels. Depending on the energy regime in which the scattering occurs, a different expression for the cross section will come into play, with fewer or more systematic errors, or the process will be affected by certain backgrounds.

Multivariate techniques applied to classification are methods used to categorize elements into different groups or classes, taking into account multiple variables or features. Some of the most commonly used techniques in this regard are Discriminant Analysis and Logistic Regression. Both techniques seek to find a linear combination of variables that maximizes the separation between different classes. They use the information from the input variables to construct a discriminant function that allows for the classification of new observations. The difference lies in the initial assumptions required about the explanatory variables, with logistic regression being much more versatile in this regard. The choice of the appropriate technique depends on the type of data, the complexity of the problem, and the specific objectives of the analysis. It is important to consider that each technique has its own advantages and limitations, so it should be carefully selected based on the characteristics of the problem and the available data.

In this Bachelor's thesis, starting with a general discussion on the issues of statistical classification, the formal and methodological aspects of different multivariate techniques are examined based on their suitability for the nature of the input variables. These techniques are applied to the classification of final state topologies for the SBN experiment.

Índice

1 Introducción **5**

2 Física de Neutrinos **7**

2.1 Oscilaciones de Neutrinos . . . 7

2.1.1 Evidencia experimental . . . 7

2.1.2 Descripción física . . . 8

2.1.3 Oscilación de dos y tres sabores . . . 8

2.1.4 Anomalías y neutrino estéril . . . 11

2.2 Experimento SBN . . . 13

2.2.1 Detectores y tecnología LAr-TPC . . . 13

2.2.2 Scattering neutrino-argón . . . 14

2.2.3 Separación de Topologías . . . 15

3 Técnicas de Aprendizaje Supervisado **17**

3.1 Análisis Discriminante . . . 17

3.1.1 Clasificación bayesiana . . . 17

3.1.2 Análisis discriminante lineal y cuadrático . . . 18

3.2 Regresión Logística . . . 19

3.2.1 Regresión logística simple y múltiple . . . 19

3.2.2 Estimación de los parámetros por máxima verosimilitud . . . 21

3.2.3 Regresión logística multinomial . . . 23

3.3 K-Nearest Neighbours . . . 24

3.4 Métricas para evaluación de los modelos . . . 25

4 Aplicación a Topologías de Estado Final **27**

4.1 Datos, variables y clases de respuesta . . . 27

4.2 Análisis Exploratorio Univariante . . . 27

4.3 Análisis Bivariante y Punto de Partida . . . 29

4.3.1 Separación de dos variables para 2 clases . . . 30

4.3.2 Separación de dos variables para 4 clases . . . 32

4.4 Resultados del Aprendizaje Supervisado . . . 34

4.4.1 Análisis Discriminante . . . 34

4.4.2 Regresión Logística . . . 35

4.4.3 K-Nearest Neighbours . . . 39

5 Conclusiones **41**

1 Introducción

El interés en este proyecto surge de mi motivación por adentrarme en el mundo de la Inteligencia Artificial y el Machine Learning, como consecuencia de los grandes avances que se están produciendo en este área y su implementación en diversos ámbitos. Con la idea de poder utilizar los conocimientos aprendidos en este Doble Grado, el tema planteado compatibiliza las dos perspectivas, de tal manera que incluye un planteamiento formal de los contenidos teóricos propios de la Matemática y la Física necesarios para abordar un problema, a la vez que desarrolla los modelos predictivos que permitirán llevar a cabo su resolución.

Más en concreto, en los experimentos de la Física de Partículas, la gran cantidad de sucesos registrados hacen de ésta un área idónea para la aplicación de modelos de Aprendizaje Automático. Estos modelos permiten un aprendizaje previo basado en un etiquetado de los datos, que permite tomar decisiones o hacer predicciones en función de un conjunto de variables explicativas. Este trabajo se contextualiza en el marco de la detección y medición del fenómeno de oscilaciones de neutrinos, en particular, dentro del **experimento *SBN*** para la búsqueda del hipotético **neutrino estéril**, en el Fermi National Accelerator Laboratory (*Fermilab*). Se trata de un tema candente en la investigación del momento con núcleo en uno de los centros de investigación más punteros del mundo.

Los neutrinos son partículas elementales de masa muy pequeña que interaccionan débilmente. Su confirmación experimental llegó a mediados del siglo pasado y, desde entonces, multitud de experimentos dedicados han sido puestos en marcha con el fin de determinar sus características y descubrir los mecanismos que rigen su naturaleza a nivel fundamental. En esto, el experimento SBN, cuyo nombre procede de las siglas de *Short-Baseline Neutrino program* o, en español, programa de Neutrinos de Corto Recorrido, es un ambicioso proyecto que surge con el objetivo, entre otros, de resolver ciertas inconsistencias entre mediciones de experimentos anteriores y así decantar la balanza hacia la existencia o no de un nuevo tipo de neutrino. Gracias a la tecnología de las cámaras de deriva de argón líquido, ***Liquid Argon Time Projection Chambers*** o ***LAr-TPCs***, se recogerán millones de sucesos de scattering de neutrinos en núcleos de argón, obteniéndose un alto nivel de significación estadística en las mediciones.

La importancia de la clasificación de las **topologías de estado final**, es decir, qué partículas se observan tras la dispersión con el núcleo y qué se puede decir de ellas, recae en la utilidad de tener sucesos categorizados a la hora de realizar análisis estadístico. Diferentes topologías están mejor explicadas mediante procesos físicos particulares, que pueden conocerse tanto a nivel teórico como experimental con distinta precisión. Dependiendo del régimen energético en el que ocurra el scattering entrará en juego una u otra expresión para la sección eficaz, pocos o muchos errores sistemáticos o el proceso estará afectado por una serie de backgrounds determinados. Teniendo esto en cuenta, parece ser que si se consigue extraer lo mejor de los datos y tener, en consecuencia, una mejor separación de sucesos, es posible un análisis pormenorizado del cual obtener determinaciones más precisas de los parámetros de interés. En última instancia, ello conduce a la confirmación o desestimación de las hipótesis bajo estudio.

Una vez fijada la motivación y el interés de nuestro objetivo, pasemos a ver cómo está estructurada esta memoria. En cada uno de los capítulos se remarcan los elementos abordados, de qué manera se han estudiado y las principales referencias utilizadas.

En el Capítulo 2, se comienza por familiarizarse con el fenómeno de las oscilaciones, para lo cual la principal fuente de consulta es [1]. Después, se presenta el proyecto SBN y sus principales elementos, destacando la importancia de las LAr-TPCs y su funcionamiento. Para ello se utiliza el informe en [2].

A partir de una discusión general sobre la problemática de la clasificación estadística, en el Capítulo 3, se proponen diferentes modelos según su adecuación al carácter de las variables de entrada y se trata de dar una visión completa y actualizada de sus aspectos fundamentales, su implementación y su validación. Las técnicas multivariantes elegidas son el **Análisis Discriminante** y la **Regresión Logística** y se desarrollan siguiendo las referencias [3], [4] y [5].

Finalmente, en el Capítulo 4, la clasificación estadística se aplica a la separación de topologías de estado final en el experimento SBND. Se comienza llevando a cabo un extenso análisis exploratorio e introduciendo un método clasificación muy básico pero intuitivo. Después, se muestran sistemáticamente los resultados de los modelos y se combinan con las interpretaciones físicas convenientes. La manera de presentar los resultados se inspira en [6]. Por su parte, los procedimientos computacionales y gráficos relacionados se apoyan principalmente en [7], [8] y [9].

2 Física de Neutrinos

2.1 Oscilaciones de Neutrinos

A diferencia de los leptones cargados, que pueden ser detectados cuando ionizan los átomos del medio que atraviesan, los neutrinos nunca se detectan directamente, sino a través de sus interacciones débiles. El sabor de un neutrino se define como el del leptón cargado junto al que dicho neutrino se genera en una corriente cargada (CC), como la de la Figura 1.

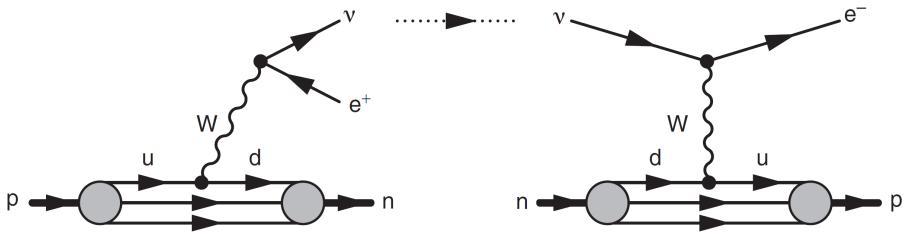

Figura 1: Producción de un neutrino/antineutrino en CC con leptón asociado e^+/e^-. Se trata por tanto de un $\nu_e/\overline{\nu}_e$ [1].

Por mucho tiempo se creyó que los neutrinos ν_e, ν_μ y ν_τ eran partículas sin masa y que el número leptónico se conservaba familia a familia[a]. No obstante, el descubrimiento de las oscilaciones, es decir, la posibilidad de que un neutrino creado con un sabor dado sea más tarde medido con un sabor diferente, requiere que las masas de estas partículas sean no nulas y que el número leptónico por familia se viole de manera flagrante en el proceso. La confirmación de este fenómeno deja la puerta abierta a nueva física.

2.1.1 Evidencia experimental

Aunque a finales de los años 90 se sabía relativamente poco sobre los neutrinos, diversas fuentes experimentales habían reportado inconsistencias en las mediciones de los llamados neutrinos solares. En el Sol, la fusión nuclear produce un gran flujo de neutrinos del electrón, y no de otros sabores, que puede llegar a ser detectado en nuestro planeta. Midiendo en diferentes regiones del espectro energético, los experimentos de *Homestake* primero, y *GALLEX* y *SAGE* después, detectaron un déficit de estos ν_e solares. Ello sería confirmado con la publicación de los datos del experimento *Super-Kamiokande* [10], diseñado más específicamente para este propósito. No obstante, todos estos experimentos estaban dedicados principal o exclusivamente a la detección del ν_e y poco se podía decir sobre los ν_μ o ν_τ. De esta forma, era imposible saber si dicho déficit era resultado de una inconsistencia teórica en cuanto al flujo total de neutrinos o, por el contrario, de algún otro fenómeno.

Con el fin de solventar esta cuestión se construyó el Observatorio de Neutrinos de Sudbury (*SNO*), cuya novedad era la de detectar neutrinos solares a través de tres pro-

[a]El número leptónico es el número cuántico resultante de la resta de leptones y antileptones en un proceso de partículas elementales. Su violación familia a familia indicaría que hay procesos permitidos en los que el número leptónico no se conserva para alguna de las tres familias, pero sí en total, como ocurriría en la desintegración hipotética $\mu^- \to e^- + \gamma$. En realidad, la desintegración que se produce es $\mu^- \to e^- + \overline{\nu}_e + \nu_\mu$, que lo conserva familia a familia.

cesos físicos, cada uno con diferente sensibilidad a los flujos de cada uno de los sabores del neutrino. Así, se demostró que el flujo total de neutrinos solares concordaba con lo esperado teóricamente, pero que en lugar de estar conformado únicamente por ν_e, este tiene una gran parte de ν_μ y ν_τ. Puesto que ni los ν_μ ni los ν_τ se generan en la fusión solar, resulta evidente que los neutrinos cambian de sabor en su propagación por el espacio.

2.1.2 Descripción física

Las transformaciones de sabor observadas en *SNO* [11] y demás experimentos se explican gracias al fenómeno de las oscilaciones de neutrinos. Para poder introducir las expresiones que dictan las probabilidades de dichas oscilaciones, es fundamental primero establecer la distinción entre autoestados de masa y autoestados débiles o de sabor cuando se habla de neutrinos. En física de partículas, lo que referimos como estados físicos, o partículas fundamentales, son los estados estacionarios del hamiltoniano para partícula libre, también llamados autoestados de masa. Para los neutrinos, estos autoestados se notan ν_1, ν_2 y ν_3 y no se corresponden con aquellos autoestados de sabor, que son el resultado, junto con su correspondiente leptón cargado, de una interacción débil. Ambas bases de neutrinos se relacionan mediante una matriz unitaria U como

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix}. \tag{2.1}$$

Los estados de sabor son aquellos que se crean en los vértices de interacción débil y realmente los que viajan son estados de masa. Entonces, necesariamente, los primeros deben estar descritos por una superposición lineal coherente de los segundos en su travesía por el espacio, hasta que su función de onda colapse. Como se ilustra en la Figura 2 para un positrón, cualquiera de los tres autoestados de masa se puede producir junto con un leptón cargado en una interacción débil. Esto lleva a definir al neutrino de sabor α como la superposición de los estados de masa que se producen junto al leptón cargado de sabor α, l_α^+, en la desintegración $W^+ \to l_\alpha^+ + \nu$, o sea,

$$|\nu_\alpha\rangle = U_{\alpha 1}^* |\nu_1\rangle + U_{\alpha 2}^* |\nu_2\rangle + U_{\alpha 3}^* |\nu_3\rangle, \qquad \alpha = e, \mu, \tau, \tag{2.2}$$

donde los números $U_{\alpha i}^*$ son las amplitudes de la desintegración leptónica de W^+ y son elementos de U^*, la matriz transpuesta conjugada de U.

2.1.3 Oscilación de dos y tres sabores

Si las masas de los estados ν_1, ν_2 y ν_3 no son iguales, habrá diferencias de fase entre las distintas componentes de la función de onda y, consecuentemente, se darán las oscilaciones. Para ilustrar este fenómeno y sus implicaciones físicas, es suficiente considerar dos sabores, sean ν_e y ν_μ, como superposición de los estados ν_1 y ν_2. La interpretación obtenida es extensible a tres sabores, añadiendo algunas particularidades.

Considérense, pues, los autoestados de masa propagándose como ondas planas de la forma

$$|\nu_1(t)\rangle = |\nu_1\rangle e^{-ip_1 \cdot x}, \qquad |\nu_2(t)\rangle = |\nu_2\rangle e^{-ip_2 \cdot x}. \tag{2.3}$$

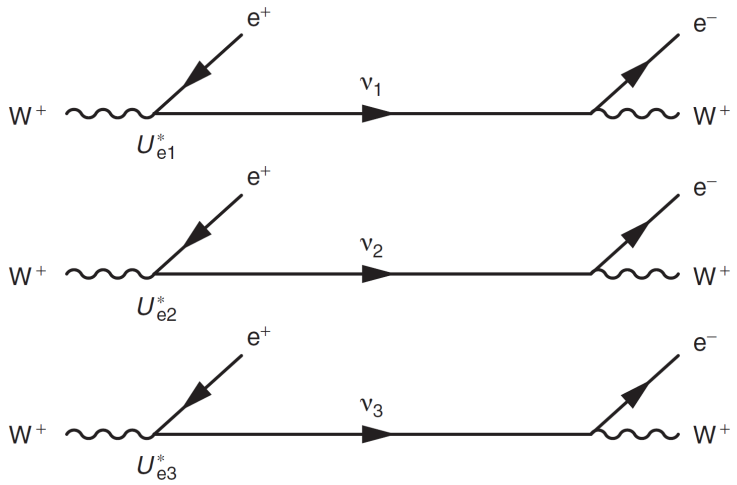

Figura 2: Vértice de interacción $We^+\nu_e$ en términos de los autoestados de masa. [1]

Para dos sabores la matriz de transformación unitaria se puede expresar en función de un único parámetro θ,

$$\begin{pmatrix} \nu_e \\ \nu_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}, \tag{2.4}$$

de forma que un estado en $t = T$ y $x = L$ se escribe

$$|\psi\,(L,T)\rangle = \cos\theta\,|\nu_1\rangle\,e^{-i\phi_1} + \sin\theta\,|\nu_2\rangle\,e^{-i\phi_2}, \tag{2.5}$$

donde las fases son $\phi_i = p_i \cdot x = E_i T - p_i L$.

Se aplica ahora la transformación inversa de 2.4 para escribir 2.5 en función de los estados de sabor, queriendo obtener una expresión del tipo

$$|\psi\,(L,T)\rangle = c_e\,|\nu_e\rangle + c_\mu\,|\nu_\mu\rangle\,. \tag{2.6}$$

Y, desarrollando, la probabilidad de que un neutrino, originariamente ν_e, produzca un muon al interaccionar, o sea, oscile de ν_e a ν_μ, es

$$P\left(\nu_e \to \nu_\mu\right) = c_\mu c_\mu^* = \sin^2(2\theta)\sin^2\left(\frac{\Delta\phi_{12}}{2}\right) \tag{2.7}$$

con $\Delta\phi_{12} = \phi_1 - \phi_2 = (E_1 - E_2)\,T - (p_1 - p_2)\,L$.

Luego la probabilidad de oscilación depende del llamado *mixing angle* θ y de la diferencia de fase entre los autoestados de masa, que puede ser reescrita en función de la diferencia de las masas de ν_1 y ν_2 al cuadrado, el *mass splitting* Δm^2_{21}, aproximadamente como

$$\Delta\phi_{12} \approx \frac{m_2^2 - m_1^2}{2p} L = \frac{\Delta m^2_{21}}{2p} L. \tag{2.8}$$

Finalmente, la probabilidad de oscilación se expresa

$$P\left(\nu_e \to \nu_\mu\right) = \sin^2\left(2\theta\right)\sin^2\left(\frac{\Delta m^2_{21} L}{4E_\nu}\right), \tag{2.9}$$

siendo su complementaria la probabilidad de supervivencia, $P\left(\nu_e \to \nu_e\right)$, como se observa en la Figura 3.

La ecuación 2.9 es fundamental. La amplitud de la oscilación se rige por $\sin^2\left(2\theta\right)$. Además, para poder oscilar, los neutrinos han de tener masas no nulas y estas ser diferentes entre sí. Si Δm^2 toma valores pequeños, las oscilaciones solo se dan a distancias muy lejanas. Entonces, fijada una energía para el haz de neutrinos, dar valores a la *baseline* L permite explorar diferentes regiones del proceso y, contrastando con las observaciones experimentales, extraer cotas para los *mixing angles* y *mass splittings*. Es por ello que en los últimos años han aparecido multitud de detectores, tanto de base larga como de base corta, *long-baseline* y *short-baseline*, con el fin de restringir cada vez más estos valores.

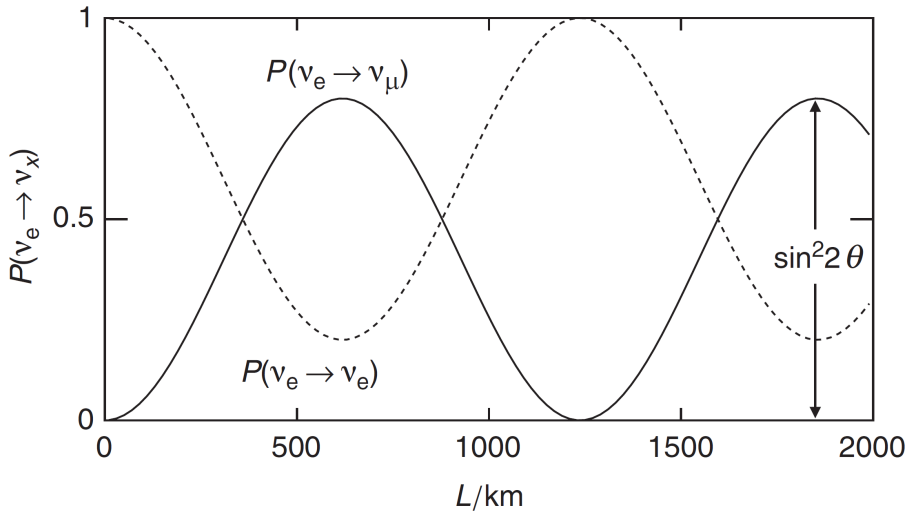

Figura 3: Probabilidad de oscilación $P\left(\nu_e \to \nu_\mu\right)$ y probabilidad de supervivencia $P\left(\nu_e \to \nu_e\right)$ para dos sabores en función de L para $E_\nu = 1\text{GeV}$, $\Delta m^2 = 0.002\text{eV}^2$ y $\sin^2\left(2\theta\right) = 0.8$ [1].

Notas para tres sabores En el muy avalado paradigma de tres neutrinos, la relación entre las dos bases de autoestados está descrita en 2.1 por la matriz unitaria de Pontecorvo-Maki-Nakagawa-Sakata (PMNS) o matriz de mezcla, cuyos elementos constituyen los parámetros fundamentales del sector fermiónico del Modelo Estándar. Esta matriz es profundamente no diagonal, o sea, los neutrinos se mezclan fuertemente, mucho más de lo que lo hacen los sabores del quark con la matriz CKM[b]. Se puede demostrar que las nueve entradas complejas de dicha matriz son reducibles a solo cuatro parámetros, tres *mixing angles* reales θ_{12}, θ_{13}, θ_{23} y una fase δ_{CP}[c]. En este caso también son tres los *mass*

[b]Contiene información sobre las desintegraciones débiles que cambian el sabor de los quarks.

[c]El parámetro δ_{CP} está relacionado con la rotura de la simetría CP, necesaria para explicar el exceso de materia sobre antimateria en el universo. Si existen tres familias de leptones es porque, con solo dos, esta fase es absorbida y la matriz de mezcla es real, no habiendo cabida para la asimetría.

splittings, de los cuales solo dos son independientes. Según resultados actuales[d] [12]:

$$|U| \sim \begin{pmatrix} 0.85 & 0.50 & 0.17 \\ 0.35 & 0.60 & 0.70 \\ 0.35 & 0.60 & 0.70 \end{pmatrix},$$

$$\begin{aligned} \sin^2(2\theta_{12}) &= 0.87 \pm 0.04, \\ \sin^2(2\theta_{23}) &> 0.92, \\ \sin^2(2\theta_{13}) &\approx 0.10 \pm 0.01, \end{aligned} \tag{2.10}$$

$$\begin{aligned} \Delta m^2_{21} = m^2_2 - m^2_1 &= (7.6 \pm 0.2) \times 10^{-5}\text{eV}^2, \\ \left|\Delta m^2_{32}\right| = \left|m^2_3 - m^2_2\right| &= (2.3 \pm 0.1) \times 10^{-3}\text{eV}^2. \end{aligned}$$

2.1.4 Anomalías y neutrino estéril

Son muchos los experimentos que han constatado y caracterizado el fenómeno de las oscilaciones para tres neutrinos. Sin embargo, durante las últimas dos décadas, numerosos resultados, agrupados bajo el nombre de *anomalías de base corta*, parecen indicar la existencia de un nuevo neutrino que no interaccionaría débilmente, el llamado *neutrino estéril*. Esta partícula, teorizada por Pontecorvo en 1967, debe ser observada indirectamente a través de la mezcla con los neutrinos del Modelo Estándar y afectaría a las oscilaciones entre los tres sabores que sí interaccionan.

Una distinción que se suele hacer en función de lo que se mide en el detector es la de experimentos de *aparición* de leptones de un sabor que no se espera o, por el contrario, de *desaparición* del leptón indicado, donde se miden menos de los predichos. En lo que concierne a las anomalías de base corta, los datos experimentales pueden dividirse en tres canales: aparición del ν_e, desaparición del ν_e y desaparición del ν_μ. La más importante de ellas tuvo su origen en el *Liquid Scintillator Neutrino Detector (LSND)*. Allí se midió un exceso de ν_e que más tarde sería confirmado por *MiniBooNE* [13].

En el formalismo de 3+1 sabores,

$$\begin{pmatrix} \nu_e \\ \nu_\mu \\ \nu_\tau \\ \nu_s \end{pmatrix} = \begin{pmatrix} U_{e1} & U_{e2} & U_{e3} & U_{e4} \\ U_{\mu 1} & U_{\mu 2} & U_{\mu 3} & U_{\mu 4} \\ U_{\tau 1} & U_{\tau 2} & U_{\tau 3} & U_{\tau 4} \\ U_{s1} & U_{s2} & U_{s3} & U_{s4} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \\ \nu_4 \end{pmatrix}, \tag{2.11}$$

suponiendo que $\Delta m^2_{41} \gg \left|\Delta m^2_{32}\right|, \Delta m^2_{31}$, las oscilaciones a corta distancia están suficientemente bien descritas por la fórmula de oscilación para dos sabores (véase 2.9):

$$P\left(\nu_\alpha \to \nu_\beta\right) = \sin^2\left(2\theta_{\alpha\beta}\right) \sin^2\left(\frac{\Delta m^2_{41} L}{4E}\right). \tag{2.12}$$

Cada canal de oscilación queda determinado por un *mixing angle* efectivo $\theta_{\alpha\beta}$, esto es,

[d] Actualmente existe un problema en la jerarquía de masas de los neutrinos y es que no se puede distinguir si $m_3 > m_2$ o si bien $m_2 > m_3$. Por ello, $\left|\Delta m^2_{32}\right|$ en valor absoluto.

$$
\begin{aligned}
\nu_\mu \to \nu_e : \sin^2\left(2\theta_{\mu e}\right) &\equiv 4\left|U_{\mu 4}\right|^2\left|U_{e4}\right|^2, \\
\nu_e \to \nu_e : \sin^2\left(2\theta_{ee}\right) &\equiv 4\left|U_{e4}\right|^2\left(1-\left|U_{e4}\right|^2\right), \\
\nu_\mu \to \nu_\mu : \sin^2\left(2\theta_{\mu\mu}\right) &\equiv 4\left|U_{\mu 4}\right|^2\left(1-\left|U_{\mu 4}\right|^2\right).
\end{aligned}
\tag{2.13}
$$

Según 2.13, que la transición $\nu_\mu \to \nu_e$ ocurra implica necesariamente que las probabilidades de supervivencia de e y μ sean estrictamente menor que uno. Esta dependencia puede ser usada para restringir el espacio paramétrico si se miden conjuntamente apariciones ν_e y desapariciones ν_e y ν_μ. La Figura 4 es el resultado del análisis que tiene en cuenta todos los datos de aparición y desaparición de una multitud de experimentos. En ella se puede observar total incompatibilidad de las regiones permitidas para los parámetros en los diferentes canales.

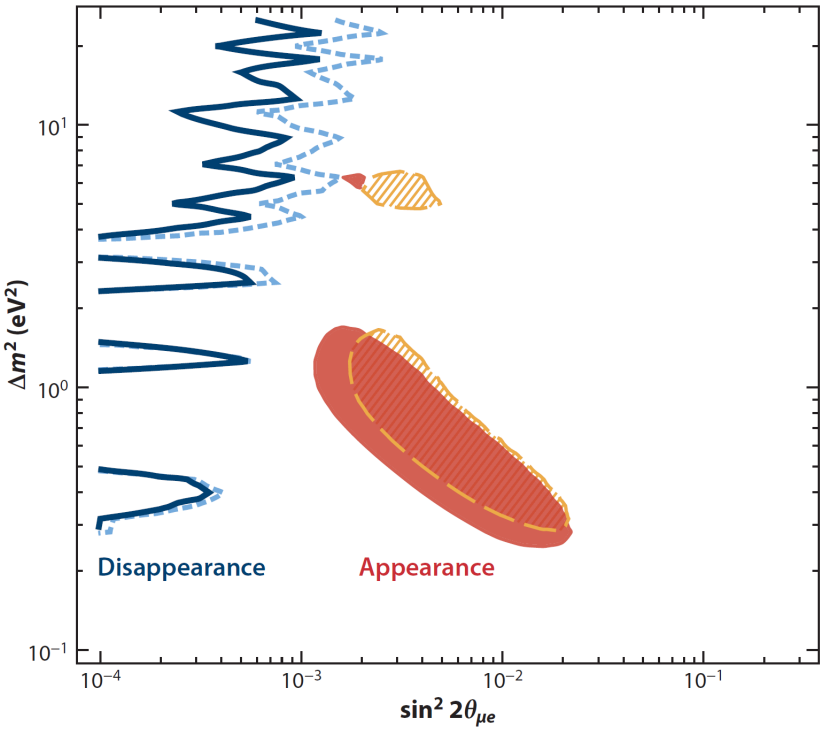

Figura 4: En el plano $\sin^2\left(2\theta_{\mu e}\right)$ frente a Δm^2, regiones permitidas para desaparición, a la izquierda de las líneas en azul, y aparición, en el interior de las regiones en rojo. [2]

Se hace así evidente la complejidad del panorama experimental del neutrino estéril. Mientras que los datos apoyando su existencia y la mezcla con los sabores activos esperan confirmación, varios experimentos que deberían estar detectando dicha mezcla no han observado nada fuera del paradigma de tres sabores. Por tanto, resulta primordial investigar si realmente la teoría del neutrino estéril es compatible con las anomalías. Para ello, una nueva gama de experimentos de precisión es estrictamente necesaria, y es aquí donde el *Short-Baseline Neutrino (SBN) program* entra en juego.

2.2 Experimento SBN

El proyecto SBN [14] en Fermilab viene a resolver el problema con las anomalías de base corta, que pueden estar escondiendo la existencia de un nuevo neutrino, que solo interacciona gravitacionalmente y que es más pesado que sus semejantes, del orden del eV. SBN es muy sensible en los canales de aparición $\nu_\mu \rightarrow \nu_e$ y desaparición $\nu_\mu \rightarrow \nu_\mu$, atajando directamente las anomalías observadas en LSND y MiniBooNE y teniendo un impacto decisivo en las discrepancias entre los diferentes conjuntos de datos. Ya sea con el descubrimiento de una nueva partícula o con una explicación a las anomalías desde dentro del Modelo Estándar, SBN acabará dando la respuesta definitiva al eterno rompecabezas de las oscilaciones y esclareciendo, en parte, el panorama de la física de neutrinos.

2.2.1 Detectores y tecnología LAr-TPC

Los experimentos con haces de neutrinos suelen incorporar dos detectores, uno suficientemente cerca de la fuente como para medir el espectro de energía de los neutrinos previo a oscilación alguna, y otro lejos para volver a medirlo tras las oscilaciones. El uso combinado de ambos hace que muchos errores sistemáticos se cancelen, permitiendo medidas de gran exactitud. Como esquematiza la Figura 5, en SBN, este papel lo asumen el *Short-Baseline Near Detector (SBND)* y el *ICARUS far detector*, situados a 110 m y 600 m de la fuente, respectivamente. Además, entre ellos, SBN emplea un tercer detector, *MicroBooNE* (no confundir con MiniBooNE). Su acción conjunta permitirá comprobar los parámetros de las anomalías pasadas a un nivel de significación $\geq 5\sigma$.

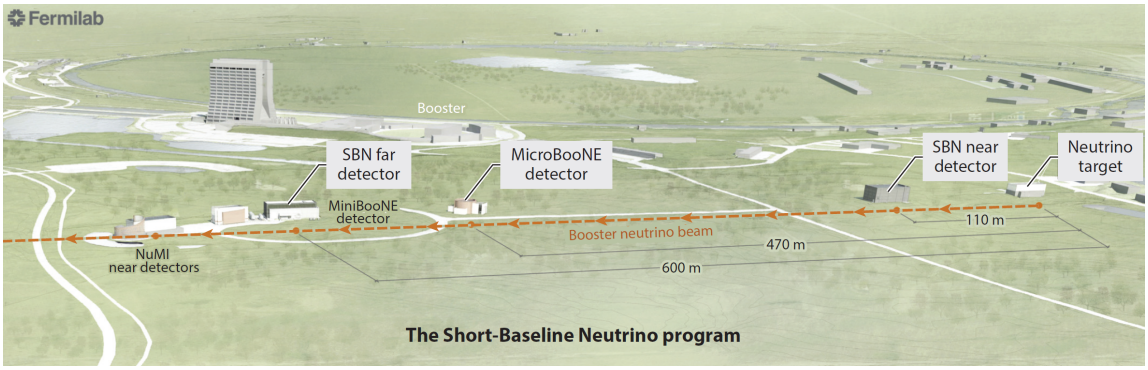

Figura 5: Vista aérea de SBN. Desde la derecha, el haz sigue la línea discontinua naranja en su encuentro con los tres detectores principales del proyecto, cuyas posiciones están indicadas [2].

Los tres detectores emplean cámaras de deriva de argón líquido, *Liquid Argon Time Projection Chambers (LAr-TPCs)*[15], que proporcionan una información muy granular y detallada acerca de las partículas en el estado final, esto es, tras la interacción de los neutrinos con los núcleos de argón. Situadas a lo largo de un mismo haz, el *Booster Neutrino Beam (BNB)*, todas funcionan de manera idéntica y solo se diferencian en su tamaño. Estando a diferente distancia de la fuente, es preciso que la más lejana, la de ICARUS, sea más grande. De esta forma, aunque el haz llegue menos colimado, más eventos son recogidos. El hecho de que los eventos en los detectores cercano y lejano estén fuertemente correlacionados se reporta en una cancelación significativa de los errores sistemáticos

referentes al flujo de neutrinos y a la sección eficaz de la interacción neutrino-núcleo, los dos mayores obstáculos en los experimentos de oscilaciones.

En cuanto al BNB, el haz de neutrinos se crea al hacer colisionar protones a 8 GeV sobre berilio. Entonces, se genera un haz hadrónico secundario, principalmente compuesto por piones. A continuación, los mesones se desintegran produciendo neutrinos del muon y del electrón. Más concretamente, BNB tiene dos modos, neutrino y antineutrino, dominados por ν_μ y $\overline{\nu}_\mu$ en torno al 93 %, respectivamente. Hay que aclarar que tanto la composición del haz como las conclusiones sobre los errores sistemáticos o sobre la sensibilidad del experimento no se extraen aún de los datos tomados por SBN, si no de anteriores medidas de MiniBooNE, que lleva funcionando más de 15 años, y a través de potentes herramientas de simulación.

LAr-TPC Una cámara de deriva de argón líquido consta de un gran volumen de argón líquido ultrapuro rodeado por una superficie que actúa como cátodo de alta tensión en una cara y una superficie que hace de ánodo opuesta. La Figura 6 muestra su funcionamiento. Cuando un neutrino experimenta una corriente cargada o neutra con un núcleo de argón, las partículas cargadas resultantes ionizan y excitan los átomos de argón a medida que se propagan en el líquido. Los electrones liberados son arrastrados en el medio bajo la influencia del campo eléctrico entre el cátodo y el ánodo, generando pequeñas corrientes en hilos de detección ubicados en el lado del ánodo. Dichos hilos están estrechamente espaciados formando planos y, para generar imágenes bi o tridimensionales de las trayectorias de las partículas, se utilizan dos o tres de estos, orientados en diferentes ángulos. En otras palabras, el detector es un calorímetro de muestreo fino, totalmente activo, con capacidades de seguimiento de partículas a nivel de milímetros.

Los átomos de argón ionizados y excitados centellean dando una señal crítica para el funcionamiento de las TPCs, al proporcionar el t_0 de la interacción dentro de la cámara o, equivalentemente, la posición a lo largo de la dirección de arrastre. Ello se debe a que la velocidad de arrastre de los electrones en la cámara es pequeña, alrededor de $1.6\, mm\, \mu s^{-1}$, y es entonces el fotón de desexcitación el que marca el comienzo del suceso, pues alcanza mucho antes el detector.

Una ventaja directa de las LAr-TPCs sobre los detectores Cherenkov, presentes en MiniBooNE, es la capacidad para distinguir electrones de fotones. Ello permitirá reducir notablemente los *backgrounds* en el exceso de ν_e observado por MiniBooNE. Por otra parte, la tecnología de las TCPs está siendo clave en la mejora de la selección y reconstrucción de los eventos, uno de los focos actuales del programa.

2.2.2 Scattering neutrino-argón

El conocimiento de las interacciones neutrino-núcleo es fundamental para afinar las medidas en los experimentos de oscilaciones. Mediante el estudio de millones de ellas, SBN tiene el potencial para transformar nuestro entendimiento de las mismas al reducir la incertidumbre estadística a un umbral sin precedentes. Como detector cercano, será SBND el que observe el mayor flujo de neutrinos, registrando más de dos millones de interacciones por año[e].

[e] Asumiendo 2.2×10^{20} POT. *Proton On Target* es una medida de rendimiento en un experimento de blanco fijo, así como lo es la luminosidad en un acelerador de haz contra haz.

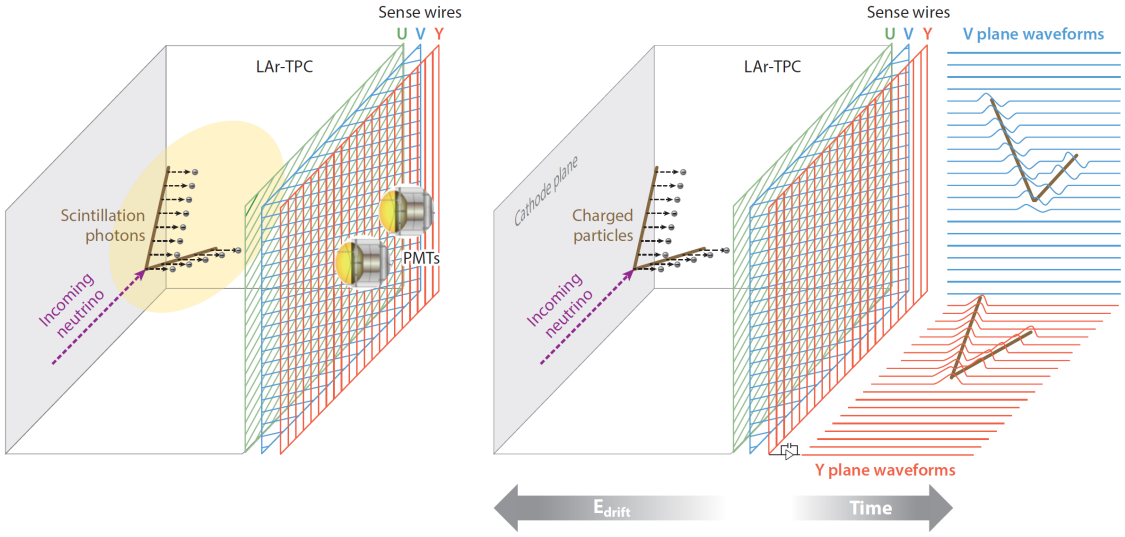

Figura 6: Principio de funcionamiento de una LAr-TPC. [2]

Dada la composición del haz BNB, la mayoría de los sucesos serán corrientes cargadas (CC) con ν_μ, mientras que las corrientes neutras (NC), aproximadamente un tercio de los eventos, no son útiles en oscilaciones al no identificar el sabor del neutrino incidente. También habrá una minoría de sucesos del ν_e que son de importante consideración. Por una parte, constituyen lo que se llama un background irreducible, es decir, una contaminación presente per se en el haz que es necesario no se confunda con ν_μ que hayan oscilado. Así, lo que se debe cuantificar como oscilaciones $\nu_\mu \rightarrow \nu_e$ son solo aquellas que se midan por encima de tal background. Por otra, la propia existencia de estos sucesos con ν_e permite hacer estadística sobre sus interacciones, como un canal independiente al anterior.

En la Figura 7, además del número total de eventos para 6.6×10^{20} POT, estos se pueden ver clasificados según el número de piones en el estado final. El hecho de separar topologías de este modo es crucial y se estudiará en detalle en el apartado siguiente. Se observa que los eventos mayoritarios son esos donde no resultan piones, donde tan solo habrá un muon saliente y uno o más nucleones en retroceso. La sección eficaz de una colisión depende fuertemente tanto de las interacciones en el estado final como de efectos nucleares. Así, un tipo de evento tan limpio, sumado a semejante tamaño de muestra, va a permitir medir con gran precisión el papel de los efectos nucleares en la interacción neutrino-argón.

No es casualidad que esos eventos más limpios, donde no hay muchas partículas en el estado final, sean también los menos energéticos. Debe haber una explicación que relacione la simplicidad de una colisión con la energía de la misma.

2.2.3 Separación de Topologías

Cuando un neutrino colisiona con un núcleo, el tipo de proceso que toma lugar depende radicalmente de la energía del encuentro. Grosso modo, se pueden distinguir dos regímenes.

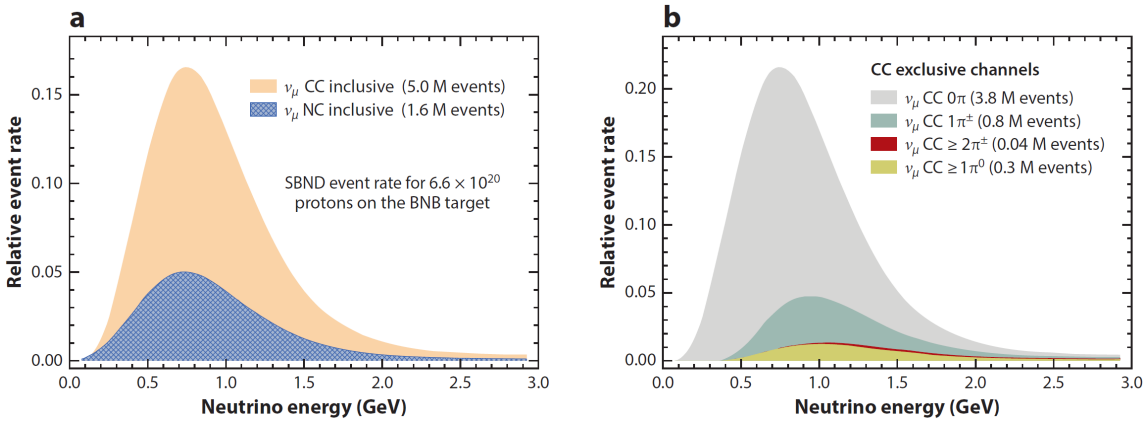

Figura 7: Eventos en SBND en 3 años de funcionamiento. (a) Espectro ν_μ CC y NC total (gráfico no apilado). (b) Espectro ν_μ CC según número de piones en el estado final (gráfico apilado). [2]

A bajas energías, se da una colisión quasi-elástica en la que el neutrino se dispersa de un nucleón, sin llegar a penetrar en su estructura interna, emitiéndose un muon. Se trata de suceso cuyo background[f] y errores sistemáticos son conocidos y están bien determinados. Esto permite una mejor reconstrucción del suceso, o sea, a partir del estado final tratar de averiguar qué había en el estado inicial y con qué energía.

A altas energías, se produce scattering fuertemente inelástico. El neutrino se adentra en la estructura interna de los nucleones, encontrándose una sopa de quarks y gluones. Entonces, tras colisionar el ν con un quark u o d, el núcleo queda destruido y sus quarks constituyentes deben agruparse de inmediato, pues su naturaleza les impide quedar libres. Es así como surgen los mesones, mayormente piones que son los más ligeros, pero también kaones. Más partículas interaccionando en el estado final conducen a una reconstrucción de peor calidad, pues hay más background y los errores sistemáticos también son más y mayores.

Esta explicación se sigue de una descripción mecano-cuántica del scattering pero, en realidad, a nivel experimental, no se puede tener la certeza sobre qué mecanismo ha llevado a un cierto estado final. No obstante, lo que se busca es agrupar eventos que, mayoritariamente, se hayan debido producir en un cierto régimen. Así, los eventos de SBND que no tienen piones en el estado final procederán en gran parte de colisiones quasi-elásticas y, según vaya aumentando el número de piones en el estado final, más inelástica habrá sido la dispersión. Y es precisamente debido a que distintos eventos albergan distintas incertidumbres, que tratarlos a todos a la vez sería desperdiciar la buenas propiedades de los unos al mezclarlos con otros cuyas variables se conocen con peor resolución. Esta es la idea en la que se basa el interés por separar topologías: misma cantidad de sucesos, si correctamente diferenciados, aportan mayor significación estadística y conducen a un análisis más robusto.

[f] En Física de Partículas, el concepto de background o ruido hace referencia a todo aquella señal que puede confundirse con la señal de interés pero que no lo es, es decir, otros sucesos que se parezcan al que se quiere medir y se confundan con este, induciendo mediciones poco precisas.

3 Técnicas de Aprendizaje Supervisado

Con una clara motivación para querer separar topologías de estado final, hay que elegir qué métodos son los apropiados para desempeñar esta tarea. Sabiendo que se va a trabajar con un conjunto de datos etiquetados en una serie de niveles o clases, dentro del Machine Learning, las técnicas a utilizar se sitúan dentro del área del Aprendizaje Supervisado y constan de dos etapas[16]:

Discriminación: Usar los datos ya etiquetados para construir un clasificador que los separe tanto como sea posible, siguiendo un cierto criterio.

Clasificación: Dado un nuevo conjunto de datos no etiquetados, usar el clasificador para asignar una etiqueta a cada nueva instancia, es decir, asociarla de manera única a una clase.

En la práctica, el procedimiento habitual consiste en dividir el dataset de trabajo, que está enteramente etiquetado, en dos conjuntos, de entrenamiento y de test. El conjunto de entrenamiento o training set se utiliza para construir el clasificador mientras que el test set se usa para su validación: para cada observación, se compara la clase asignada por el clasificador con la clase predefinida a la que realmente pertenece. Finalmente, se da una medida de la bondad del clasificador en función de las etiquetas asignadas correctamente.

Aclarado ya el contexto, se procede a introducir los aspectos formales del análisis discriminante y de la regresión logística, dos métodos propios de la Estadística Multivariante. Se presenta también el algoritmo de los k vecinos más cercanos.

3.1 Análisis Discriminante

El principal objetivo del análisis discriminante es encontrar regiones en las que separar observaciones multivariantes tratando de minimizar el error de clasificación. Las fronteras que definen dichas regiones son las llamadas funciones discriminantes y, dependiendo de las características de los datos, estas tomarán una u otra forma.

El enfoque aquí seguido [16] para la obtención de estas funciones, lineales y cuadráticas, sigue la regla de máxima probabilidad a posteriori, cuyo origen está en el Teorema de Bayes. Además, se supondrá que las variables explicativas se distribuyen según una normal multivariante.

3.1.1 Clasificación bayesiana

Dadas J categorías o poblaciones, la probabilidad de que una observación cualquiera proceda de la población j se expresa como

$$P\left(población\ j\right) = \pi_j$$

donde π_j se llama probabilidad a priori, pues representa una característica de la población que se presupone antes de tener observación alguna. Es habitual estimar las probabilidades a priori por las proporciones de las poblaciones en el total del conjunto de datos de trabajo, $\widehat{\pi}_j = n_j/N$.

Sea x una observación k-dimensional, su función de densidad condicionada a que pertenezca a la población j se escribe, por ahora, de modo general, como $f(x \mid j)$. De tal modo que la densidad conjunta de haber escogido la población j y además haber observado x en ella es

$$f(x, j) = \pi_j f(x \mid j). \tag{3.1}$$

El último ingrediente necesario para poder aplicar el Teorema de Bayes es la probabilidad marginal de x, sin importar la población. Naturalmente como suma sobre todas las poblaciones $j = 1, .., J$ posibles

$$f(x) = \sum_{j=1}^{J} \pi_j f(x \mid j). \tag{3.2}$$

Entonces, la probabilidad a posteriori de la población j dada una observación x es

$$f(j \mid x) = \frac{f(x, j)}{f(x)} = \frac{\pi_j f(x \mid j)}{\sum_{j=1}^{J} \pi_j f(x \mid j)}. \tag{3.3}$$

Dado que el denominador es el mismo para todas las poblaciones, maximizar 3.3 equivale a escoger $\widehat{j}$ como estimador máximo a posteriori de j si

$$\pi_{\widehat{j}} f\left(x \mid \widehat{j}\right) = \max\left\{\pi_j f(x \mid j) : j = 1, \ldots, J\right\} \tag{3.4}$$

3.1.2 Análisis discriminante lineal y cuadrático

Supóngase que cada una de las J poblaciones queda representada por una normal multivariante con vector de medias μ_j y matriz de covarianza común Σ. En tal caso, x es una observación de una mixtura gaussiana con pesos $\{\pi_j : j = 1, \cdots, J\}$, de forma que, condicionada al valor de j, la función de densidad se lee

$$f(x \mid j) = f\left(x \mid \mu_j, \Sigma\right) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\left(x - \mu_j\right)^T \Sigma^{-1}\left(x - \mu_j\right)\right). \tag{3.5}$$

En la práctica, las medias y matriz de covarianza son desconocidas y son sustituidas por sus estimadores habituales. El j que maximiza 3.4 también maximiza el logaritmo de esa cantidad,

$$\ln\left[\pi_j f\left(x \mid \mu_j, \Sigma\right)\right] = \ln \pi_j - \ln\left[(2\pi)^{-k/2} |\Sigma|^{-1/2}\right] - \frac{1}{2}\left(x - \mu_j\right)^T \Sigma^{-1}\left(x - \mu_j\right). \tag{3.6}$$

Desarrollando la forma cuadrática en x y deshaciéndose de los términos que no dependen de j, la función discriminante a maximizar queda

$$\delta_j(x) = \ln \pi_j + x^T \Sigma^{-1} \mu_j - \frac{1}{2} \mu_j^T \Sigma^{-1} \mu_j. \tag{3.7}$$

Nótese que se trata de una función lineal en x, de ahí el nombre discriminante lineal.

La cuestión por resolver es cómo se encuentran las fronteras de las regiones de clasificación. Simple, para dos poblaciones j y t, basta resolver la ecuación $\delta_j(x) = \delta_t(x)$. Ello deja la expresión siguiente:

$$x^T \Sigma^{-1} (\mu_j - \mu_t) = c(j,t) \tag{3.8}$$

donde $c(j,t)$ es una constante que no depende de x. Sustituyendo medias y varianza por sus versiones muestrales, la expresión obtenida tiene nombre y se trata de la **función discriminante de Fisher** entre las poblaciones j y t,

$$L_{jt}(x) = x^T S^{-1} \left(\widehat{\mu}_j - \widehat{\mu}_t\right). \tag{3.9}$$

Para J poblaciones habrá, a lo sumo, $J-1$ de estas fronteras.

Si se elimina ahora el supuesto de homogeneidad de varianzas,

$$f(x \mid j) = f\left(x \mid \mu_j, \Sigma_j\right), \tag{3.10}$$

los términos correspondientes en 3.6 no se simplifican y resulta entonces una función discriminante cuadrática,

$$\delta_j(x) = \ln \pi_j + \frac{1}{2} \ln \left|\Sigma_j\right| - \frac{1}{2} (x - \mu_j)^T \Sigma_j^{-1} (x - \mu_j). \tag{3.11}$$

3.2 Regresión Logística

Cuando la variable de respuesta de un modelo es categórica y binaria, un método a emplear es la regresión logística o modelo logit. Siendo dos los posibles resultados $Y = 0$ e $Y = 1$, es natural suponer que dicha respuesta condicionada a cada valor de la variable explicativa sigue una distribución binomial. En caso de ser $J > 2$ categorías, la asunción de binomial es sustituida por la de multinomial, con parámetros las probabilidades de los niveles de respuesta,

$$(Y \mid X = x) \to B(p_1(x)), \qquad (Y \mid X = x) \to M(1; p_1(x), \ldots, p_J(x)). \tag{3.12}$$

El enfoque que permite poner en práctica la regresión logística para una variable de respuesta politómica es el de tomar una categoría base, de referencia, y, para cada una de las categorías restantes establecer un modelo logit binario frente a esa.

Se presentan primero el modelo básico para una sola variable explicativa y el múltiple [4], ambos binarios y la estimación de sus parámetros por máxima verosimilitud [5]. Después, se generaliza para cubrir el caso de interés. En todos ellos, es de notar la ausencia de hipótesis sobre las variables explicativas como una ventaja de la regresión logística sobre el análisis discriminante, en caso de que los supuestos de normalidad no sean satisfechos.

3.2.1 Regresión logística simple y múltiple

El modelo de regresión logística simple es de la forma

$$p(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{3.13}$$

Se trata de una curva sigmoide estrictamente creciente si $\beta > 0$ y estrictamente decreciente si $\beta < 0$. Equivalentemente se puede expresar

$$\ln\left[\frac{p(x)}{1-p(x)}\right] = logit\left[p(x)\right] = \alpha + \beta x, \tag{3.14}$$

donde el cociente $p(x)/1-p(x)$ se conoce como ***odds*** y representa la ventaja de la respuesta 1 frente a la 0 para el valor observado de x, esto es, cuántas veces se obtiene un éxito por cada fracaso. El logaritmo de las odds es la transformación logística o *logit*.

En este modelo, la pendiente de la curva $p(x)$ no es constante y como mucho se puede decir que alrededor del punto de máxima pendiente la función varía $\beta/4$ por cada incremento unitario en x. Sin embargo, independiente de x, la interpretación más interesante del parámetro β viene ligada al concepto de ***odds ratio***, OR. Exponenciando en 3.14, el cociente de ventajas u odds ratio para dos valores de X es

$$OR(x_1, x_2) = \frac{e^{\alpha+\beta x_1}}{e^{\alpha+\beta x_2}} = e^{\beta(x_1-x_2)}. \tag{3.15}$$

De tal forma que e^{β} representa el aumento multiplicativo de las odds por cada unidad de incremento en X. Llevado a un ejemplo de la física, supóngase que en una cámara de deriva se quiere discriminar entre la aparición o no de una cierta partícula en función del número de trazas, X, observado. Claro que a mayor número de trazas más probable es que una de ellas pertenezca a la partícula de interés. Así, en este caso, la odds ratio $OR(x+1, x) = OR(\Delta X = 1)$ debe ser mayor que uno, luego β ser positivo. En otras palabras, los parámetros del modelo son de más fácil interpretación cuando se exponencian, pues entonces representan los efectos multiplicativos sobre las odds.

Si se consideran ahora k variables explicativas, la regresión logística toma la forma

$$logit\left[p(x_1, \ldots, x_k)\right] = \alpha + \sum_{i=1}^{k} \beta_i x_i. \tag{3.16}$$

En este caso lo más intuitivo es observar el efecto sobre las odds del incremento en una unidad de una de las variables explicativas cuando el resto quedan fijas, siendo que las odds quedan multiplicadas por la exponencial del coeficiente de la variable incrementada:

$$OR(\Delta X_l = 1/X_r = x_r,\ r \neq l) = e^{\beta_l}. \tag{3.17}$$

Si, por ejemplo, se quisiera calcular es el efecto sobre las odds de un incremento de dos unidades en la variable l y de uno en la variable t:

$$OR(\Delta X_l = 2, \Delta X_t = 1/X_r = x_r,\ r \neq l, t) = e^{2\beta_l} e^{\beta_t} \tag{3.18}$$

3.2.2 Estimación de los parámetros por máxima verosimilitud

Para una muestra de tamaño N, la variable de respuesta Y devuelve una sucesión de unos y ceros procedente de N pruebas de Bernuilli independientes. A cada una de ellas corresponde una determinada combinación $x_n = (x_{n0}, \ldots, x_{nk})^T$ $(n = 1, \ldots, N)$ de las k variables explicativas. En consecuencia, denotando $p_n = P(Y = 1 \mid X = x_n)$, el modelo logit muestral es

$$L_n = \ln\left[\frac{p_n}{1-p_n}\right] = \alpha + \sum_{i=1}^{k} \beta_i x_{ni} = \sum_{i=0}^{k} \beta_i x_{ni}, \tag{3.19}$$

$$p_n = \frac{\exp\left(\sum_{i=0}^{k} \beta_i x_{ni}\right)}{1 + \exp\left(\sum_{i=0}^{k} \beta_i x_{ni}\right)}, \tag{3.20}$$

donde $\beta = (\beta_0, \ldots, \beta_k)^T$ es el vector de parámetros a estimar.

Estimar estos parámetros por máxima verosimilitud significa maximizar la función de verosimilitud de las observaciones respecto de los parámetros del modelo. Dicha función es el producto de las funciones masa de probabilidad de las N Bernuilli independientes Y_n, que es

$$\prod_{n=1}^{N} p_n^{y_n} (1-p_n)^{1-y_n} = (1-p_n)^N \left(\prod_{n=1}^{N} \left(\frac{p_n}{1-p_n}\right)^{y_n}\right) \tag{3.21}$$

Equivalentemente, se maximiza la log-verosimilitud

$$\begin{aligned} L(\beta) &= N \ln(1-p_n) + \sum_{n=1}^{N} y_n \ln\left(\frac{p_n}{1-p_n}\right) \\ &= -N \ln\left(1 + \exp\left(\sum_{i=0}^{k} \beta_i x_{ni}\right)\right) + \sum_{n=1}^{N} y_n \left(\sum_{i=0}^{k} \beta_i x_{ni}\right) \\ &= -N \ln\left(1 + \exp\left(\sum_{i=0}^{k} \beta_i x_{ni}\right)\right) + \sum_{i=0}^{k} \left(\sum_{n=1}^{N} y_n x_{ni}\right) \beta_i \end{aligned} \tag{3.22}$$

que depende de las observaciones y_n solo a través de los estadísticos suficientes $\left\{\sum_{n=1}^{N} y_n x_{ni},\ i = 1, \ldots, k\right\}$.

Las ecuaciones de verosimilitud se obtienen derivando respecto a cada uno de los parámetros β_i e igualando a cero:

$$\sum_{n=1}^{N} x_{ni} (y_n - \widehat{p_n}) = 0, \qquad i = 0, \ldots, k \tag{3.23}$$

con $X_0 = 1$, $\widehat{p_n}$ el estimador máximo verosímil de p_n,

$$\widehat{p_n} = \frac{\exp\left(\sum_{i=0}^{k} \widehat{\beta}_i x_{ni}\right)}{1 + \exp\left(\sum_{i=0}^{k} \widehat{\beta}_i x_{ni}\right)} \tag{3.24}$$

y $\widehat{\beta}_i$ los estimadores de los parámetros. Como la log-verosimilitud es una función cóncava para los modelos logit, los estimadores máximo verosímiles de sus parámetros existen y son únicos, pero como las ecuaciones de verosimilitud no son lineales en los parámetros hay que emplear métodos numéricos.

Estimación iterativa vía Newton-Raphson La idea consiste en aproximar, en cada paso, el máximo de la función de interés por el de un polinomio de segundo grado que se parezca a esta en un entorno de la aproximación del paso anterior. Así pues, se comienza con el desarrollo de Taylor hasta segundo orden de la log-verosimilitud $L(\beta)$ en un entorno del valor inicial $\beta^{(0)}$,

$$L\left(\beta^{(0)}\right) + \left(D^{(0)}\right)^T \left(\beta - \beta^{(0)}\right) + \frac{1}{2}\left(\beta - \beta^{(0)}\right)^T H^{(0)} \left(\beta - \beta^{(0)}\right), \tag{3.25}$$

donde $D^{(0)}$ es el vector de derivadas primeras y $H^{(0)}$ es la matriz de derivadas segundas de $L(\beta)$ en $\beta = \beta^{(0)}$. El máximo se obtiene derivando con respecto a β e igualando a cero,

$$D^{(0)} + H^{(0)}\beta - H^{(0)}\beta^{(0)} = 0, \tag{3.26}$$

llevando a que la posición del máximo en la iteración (t) se obtiene a partir de la anterior $(t-1)$ como

$$\beta^{(t)} = \beta^{(t-1)} - \left(H^{(t-1)}\right)^{-1} D^{(t-1)}. \tag{3.27}$$

Queda, entonces, calcular las derivadas. Las de primer orden son, evaluadas en la aproximación del paso $(t-1)$,

$$D_i^{(t-1)} = \left.\frac{\partial L(\beta)}{\partial \beta_i}\right|_{\beta^{(t-1)}} = \sum_{n=1}^{N} x_{ni}\left(y_n - p_n^{(t-1)}\right), \tag{3.28}$$

donde

$$p_n^{(t-1)} = \frac{\exp\left(\sum_{i=0}^{k} \beta_i^{(t-1)} x_{ni}\right)}{1 + \exp\left(\sum_{i=0}^{k} \beta_i^{(t-1)} x_{ni}\right)}. \tag{3.29}$$

La entrada (j, i) de la matriz de derivadas segundas es

$$\begin{aligned}\frac{\partial^2 L(\beta)}{\partial\beta_j\beta_i} &= \frac{\partial}{\partial\beta_j}\left[\sum_{n=1}^{N} x_{ni}\left(y_n - p_n\right)\right] \\ &= \frac{\partial}{\partial\beta_j}\left[\sum_{n=1}^{N} x_{ni}\left(y_n - \frac{\exp\left(\sum_{i=0}^{k}\beta_i x_{ni}\right)}{1+\exp\left(\sum_{i=0}^{k}\beta_i x_{ni}\right)}\right)\right] \\ &= -\sum_{n=1}^{N} x_{ni}x_{nj}\frac{\exp\left(\sum_{i=0}^{k}\beta_i x_{ni}\right)}{\left[1+\exp\left(\sum_{i=0}^{k}\beta_i x_{ni}\right)\right]^2} \\ &= -\sum_{n=1}^{N} x_{ni}x_{nj}p_n\left(1-p_n\right)\end{aligned} \tag{3.30}$$

que evaluada en la aproximación del máximo en $(t-1)$ resulta

$$H_{ji}^{(t-1)} = \left.\frac{\partial^2 L(\beta)}{\partial\beta_j\beta_i}\right|_{\beta^{(t-1)}} = -\sum_{n=1}^{N} x_{ni}x_{nj}p_n^{(t-1)}\left(1-p_n^{(t-1)}\right). \tag{3.31}$$

Llevando estas expresiones a 3.27, el máximo en el paso (t) resulta

$$\beta^{(t)} = \beta^{(t-1)} + \left[X^T Q X\right]^{-1} X^T\left(y - p^{(t-1)}\right), \tag{3.32}$$

siendo $X = \left(x_{ij}\right)_{j=1,\ldots,k}^{i=1,\ldots,N}$ la matriz de diseño y Q la matriz diagonal

$$Q = diag\left(p_i^{(t-1)}\left(1-p_i^{(t-1)}\right)\right)_{i=1,\ldots,N}.$$

Las aproximaciones $p^{(t)}$ y $\beta^{(t)}$ convergen en pocas iteraciones a los estimadores máximo verosímiles $\widehat{p}$ y $\widehat{\beta}$, respectivamente. El algoritmo queda completamente detallado eligiendo como $\beta^{(0)}$ la estimación mínimo cuadrática de β y como criterio de parada una diferencia entre los parámetros de dos iteraciones consecutivas menor que un cierto umbral, sea,

$$\left|\beta_i^{(t)} - \beta_i^{(t-1)}\right| \leq 10^{-9} \qquad i = 0,\ldots,k.$$

3.2.3 Regresión logística multinomial

Considérese una variable de respuesta politómica Y con J categorías y, como en el caso binario, se quiere estimar la probabilidad de cada categoría de respuesta a través de un conjunto de variables explicativas. En lugar de tener que plantear un modelo logit binario para cada par de categorías, es suficiente con $J-1$ de ellos que emparejen cada categoría con una base o de referencia, denotada B. Cada uno de estos logit tiene sus propios parámetros

$$L_j(x) = \ln\left[\frac{p_j(x)}{p_B(x)}\right] = \alpha_j + \sum_{i=1}^{k}\beta_{ji}x_i, \qquad j = 1,\ldots,J-1. \tag{3.33}$$

y se trata de una regresión logística ordinaria que tiene como opciones pertenecer a la categoría j o a la categoría base B. En efecto, solo $J-1$ logits son necesarios, pues

cualquier otra transformación que involucre a un par de categorías, ambas distintas de la base, puede ser expresada en función de las transformaciones de las mismas dos con respecto a la base:

$$\ln\left[\frac{p_j(x)}{p_t(x)}\right] = L_j(x) - L_t(x) \qquad \forall j,t. \tag{3.34}$$

Si se expresa ahora el modelo para las probabilidades de respuesta, se pueden estas calcular y asignar entonces cada observación x a la clase que la maximice:

$$\begin{aligned} p_j(x) &= \frac{\exp\left(\alpha_j + \sum_{i=1}^{k}\beta_{ij}x_i\right)}{1+\sum_{j=1}^{J-1}\exp\left(\alpha_j + \sum_{i=1}^{k}\beta_{ij}x_i\right)} \qquad j = 1,\ldots,J-1, \\ p_B(x) &= \frac{1}{1+\sum_{j=1}^{J-1}\exp\left(\alpha_j + \sum_{i=1}^{k}\beta_{ij}x_i\right)}. \end{aligned} \tag{3.35}$$

La interpretación en términos de odds ratio es idéntica al caso binario múltiple: se incrementa una variable en una unidad y el resto quedan fijas

$$OR_j(\Delta X_l = 1/X_r = x_r,\, r \neq l) = e^{\beta_{lj}} \tag{3.36}$$

Este es el cociente de ventajas de la respuesta j frente a la base B (para un incremente unitario en la variable l). Si se quiere saber a cuál de las categorías favorece más este incremento basta realizar el cálculo anterior para todas ellas y comparar las odds ratio.

En cuanto a la estimación de los parámetros, el proceder es el mismo que en el caso binario. Esta vez, se tiene una función de verosimilitud producto de funciones masa de probabilidad multinomiales y a resolver iterativamente quedan $J-1$ sistemas de $k+1$ ecuaciones no lineales. Para más detalle consultar [5].

3.3 K-Nearest Neighbours

Hay una serie de algoritmos que no formulan ninguna hipótesis sobre la distribución de la variable de respuesta, los llamados no paramétricos. Suelen ser más pesados computacionalmente pero lograr mejores clasificaciones. Entre ellos, el de los k vecinos más cercanos, kNN, es posiblemente el de más fácil interpretación.

Para cada entrada x, se identifican las k observaciones del training set más cercanas a esta. Se cuentan cuántas de esas k pertenecen a cada clase y se asigna x a la mayoritaria. En caso de empate, la clase se asigna aleatoriamente entre las empatadas. Desde el punto de vista del formalismo bayesiano, de vuelta al principio de máxima probabilidad a posterior, el estimador kNN de que x pertenezca a la clase j es

$$\widehat{p_j}(x) = \max\left\{k_j/k : j = 1,\ldots,J\right\}, \tag{3.37}$$

donde J es el número de clases y k_j el número de observaciones de la clase j entre las k más cercanas. La cercanía entre dos observaciones se mide casi siempre en distancia euclídea y es fundamental que los datos estén tipificados, de forma que todas las variables tengan media cero y desviación típica la unidad.

Sin embargo, esta formulación puede ser problemática si las categorías del training set están desequilibradas [17]. En tal caso, se podría introducir unos pesos en la regla de clasificación tales que compensen el número de observaciones. Entonces, se tomará $\widehat{j}$ como estimador máximo a posteriori de j si

$$k_{\widehat{j}}/\pi_{\widehat{j}} = \max\left\{k_j/\pi_j : j = 1, \ldots, J\right\}, \tag{3.38}$$

donde π_j es la probabilidad a priori de la clase j. Otras modificaciones son también aceptables, como la de favorecer o castigar a alguna de las clases debido a otras consideraciones y no solo por su presencia relativa en el training set.

3.4 Métricas para evaluación de los modelos

La evaluación de un método de aprendizaje supervisado suele obtenerse de la denominada matriz de confusión, que contiene las predicciones frente a los valores reales. En el caso binario, estas son algunas de las medidas que se obtienen [18], todas mayores que cero y tanto mejores como cerca de uno estén:

Observado/Predicción	Positivo	Negativo
Positivo	Verdaderos positivos (TP)	Falsos negativos (FN)
Negativo	Falsos positivos (FP)	Verdaderos negativos (TN)

Precisión (ACC, *accuracy*): Tasa total de aciertos. Se calcula como el cociente entre los elementos de la diagonal y el total

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3.39}$$

Sensibilidad (TPR, *true positive rate*): Tasa de verdaderos positivos clasificados correctamente. Se calcula como el elemento de la diagonal entre la suma de su fila

$$TPR = \frac{TP}{TP + FN}. \tag{3.40}$$

En el argot de la Física de Partículas se conoce como **eficiencia**.

Índice predictivo positivo (PPV, *positive predictive value*): Cociente entre los verdaderos positivos y los clasificados como tal. Se calcula como el elemento de la diagonal entre la suma de su columna

$$PPV = \frac{TP}{TP + FP}. \tag{3.41}$$

También llamado **pureza**, dando a entender cómo de contaminada está la clasificación en un nivel dado.

Puntuación F_1 (F_1 score): Media armónica del TPR y el PPV. Es una medida más robusta y especialmente útil en caso de que falsos positivos y falsos negativos tengan igual importancia.

$$F_1 = \frac{2 \cdot TPR \cdot PPV}{TPR + PPV}. \tag{3.42}$$

En un problema multiclase, la matriz de confusión es cuadrada de orden el número de clases y las medidas de precisión se obtienen, para cada clase, según los mismos cálculos, por filas o por columnas. Después, se calcula una media de las medidas de todas las clases. Teniendo en cuenta que los niveles pueden estar descompensados, es conveniente que dicha media sea pesada con el número de observaciones por clase. De lo contrario, podrían obtenerse buenas métricas con un clasificador que asignara todas las instancias a la clase mayoritaria.

Otra métrica, especialmente buena para evaluar clasificadores multiclase con niveles descompensados, es el coeficiente kappa de Cohen. Es una modificación de la precisión y tiene en cuenta la posibilidad de que una clasificación correcta haya sido fruto del azar, P_C, entendiendo que cuantas más instancias tenga una clase más probable es que se le haya asignado una instancia por casualidad.

$$\kappa = \frac{ACC - P_C}{1 - P_C} \tag{3.43}$$

donde

$$P_C = \sum_{j=1}^{J} P_j^{obs} P_j^{pred} = \frac{1}{N^2} \sum_{j=1}^{J} \left(TP_j + FN_j\right) \left(TP_j + FP_j\right), \tag{3.44}$$

para una clasificación de N observaciones en J niveles. Valores mayores que 0 indican una clasificación mejor que al azar y por encima de 0.8 una muy buena.

4 Aplicación a Topologías de Estado Final

Todo el código en R para la elaboración de los gráficos y resultados de esta sección está disponible en mi repositorio de GitHub.

4.1 Datos, variables y clases de respuesta

Se trabaja con un dataset de topologías finales en eventos neutrino-argón. Se quieren clasificar en 4 clases en función del número de piones neutros π^0 y piones cargados $\pi^{\pm}$. En la Tabla 1 se describen los niveles de respuestas y las variables explicativas utilizadas, todas ellas de la información que recoge el detector SBND en el scattering. Nótese también que las categorías elegidas difieren de las de la Figura 7. Estas no son inamovibles y, entre las que se probaron, las mostradas aquí dieron el mejor resultado. Se aclara el significado de los niveles: *NoPi* si no hay piones en el estado final, *Pi0* si hay un solo pion neutro, *PiCh* si hay un solo pion cargado (independientemente de su carga) y *Other* si la suma de piones neutros y cargados es mayor o igual que dos.

Tabla 1: Descripción de las variables.

Variable de Respuesta	**Niveles**	**%(N)**
Cat = Topología ($N = 49900$)	(NoPi) 0π	12.7 (6348)
	(Pi0) $1\pi^0$	14.7 (7356)
	(PiCh) $1\pi^{\pm}$	24.1 (12009)
	(Other) $\geq 2\pi$	48.5 (24187)
Variables Reconstruidas Discretas	**Media**	**SD**
nShowers = C (Número de cascadas)	1.8	1.6
nTracks = T (Número de trazas)	2.7	1.5
nStubs = S (Número de stubs)	3.6	2.8
Variables Reconstruidas Continuas	**Media**	**SD**
maxShowerLength = LC (cm) (Cascada más larga)	36.1	39.6
maxTrackLength = LT (cm) (Traza más larga)	134.4	139.0
maxOpenAngle = AC (rad) (Ángulo de cascada más grande)	0.25	0.23

4.2 Análisis Exploratorio Univariante

Es clave en primer lugar estudiar qué variables pueden tener un mayor valor para la clasificación, viendo si poseen capacidad discriminatoria por sí solas. La Figura 8 es

una agrupación de histogramas que amontonan los datos de las cuatro categorías y los dividen por color. En el caso de las variables continuas se usan densidades suavizadas. Se evidencia rápidamente que las distribuciones de todas las variables se encuentran realmente solapadas, lo cual anuncia dificultades cuando se trata de discriminar.

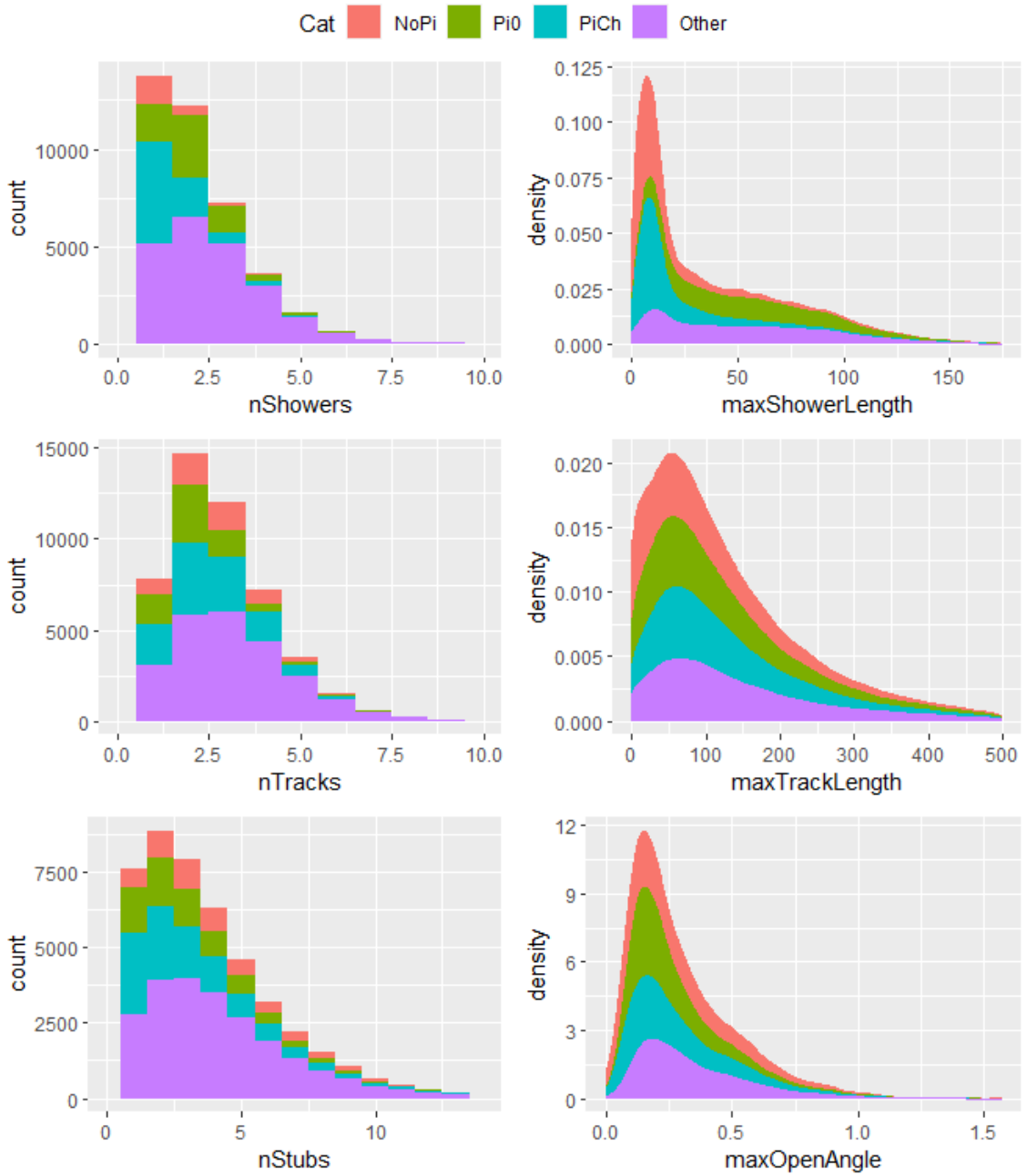

Figura 8: Histogramas y densidades suavizadas apiladas para las variables discretas y continuas, respectivamente.

Cuanto los datos no dan demasiadas pistas de primeras es importante agarrarse a otras herramientas e intentar ganar algo de intuición que ayude a atacar el problema. Con este propósito, se presentan en la Figura 9 algunos diagramas de cajas a analizar basándose en los principios físicos que subyacen. Es necesario también presentar las

medias por categoría, en la Tabla 2, ya que debido a la multitud de valores extremos no es posible apreciarlas bien en los diagramas.

Tabla 2: Medias por categoría para algunas variables. Solo a modo ilustrativo para la discusión. No se incluyen las desviaciones típicas.

	nShowers	**maxShowerLength**	**nTracks**	**maxOpenAngle**
NoPi	0.6	7.0	2.3	0.10
Pi0	2.0	52.5	2.1	0.25
PiCh	1.0	15.1	2.5	0.19
Other	2.5	49.2	3.1	0.31

Las medias del número de cascadas difieren notablemente lo cual hace pensar que va a ser buena variable de separación. La explicación es sencilla: si no hay piones no debe haber muchas partículas capaces de producir cascadas y por eso su media es cercana a cero. Cuando hay un π^0, la media debe ser cercana a dos porque este se desintegra rápidamente a dos fotones, cada uno dando lugar a una cascada. Si hay más piones, más cascadas se producirán, con lo que la media sube para la categoría *Other*. Esto se relaciona directamente con *maxShowerLength*, que es muy grande cuando hay π^0 y muy pequeña cuando no. Con respecto al número de trazas, se puede ver que la media no baja de dos unidades. En efecto, en la topología final normalmente habrá un muon relativista y un protón. Además, cada $\pi^{\pm}$ deja otra traza. Por el contrario, los π^0 no dejan traza y por eso su media se parece a la de categoría sin piones. Finalmente, y aunque las variables continuas son más difíciles de interpretar de este modo, se puede comentar que el *maxOpenAngle*, que es el ángulo de cono que contiene la cascada más abierta, debe ser mayor cuanto más energético sea el evento, o sea, mayor en los eventos con $\geq 2\pi$.

Hay que entender que esta fenomenología rara vez se observa fielmente en un evento aislado. Suceso a suceso, es difícil interpretar qué está sucediendo en el detector, pues hay algunas fuentes externas que pueden estar generando una reconstrucción imperfecta del evento. De hecho, la variable *nStubs* está relacionada con esto mismo, al considerarse como *stub* un rastro en el detector que no ha podido ser identificado claramente como cascada o como traza. En resumen, es cuando se estudian una multitud de sucesos que la fenomenología descrita debe casar con los datos.

4.3 Análisis Bivariante y Punto de Partida

Se explora la posibilidad de plantear clasificadores que usen un par de variables. Debido a que el solapamiento entre las topologías es grande, un gráfico de dispersión con todas ellas es complicado de interpretar. En la Figura 10, se prueba a incluir únicamente las dos categorías extremas, sin piones y con más de un pión, considerando solo las variables discretas. Se sigue comprobando que se trata de un ejercicio complicado, pero quizás haya algo de poder de separación si se discrimina en función del número de cascadas (*nShowers*) y el número de trazas (*nTracks*). Cuando ambos son pequeños, los puntos rojos predominan sobre los azules y, lejos del origen, aunque sigue habiendo puntos rojos, estos quedan totalmente eclipsados. Parece entonces interesante analizar esta pareja de

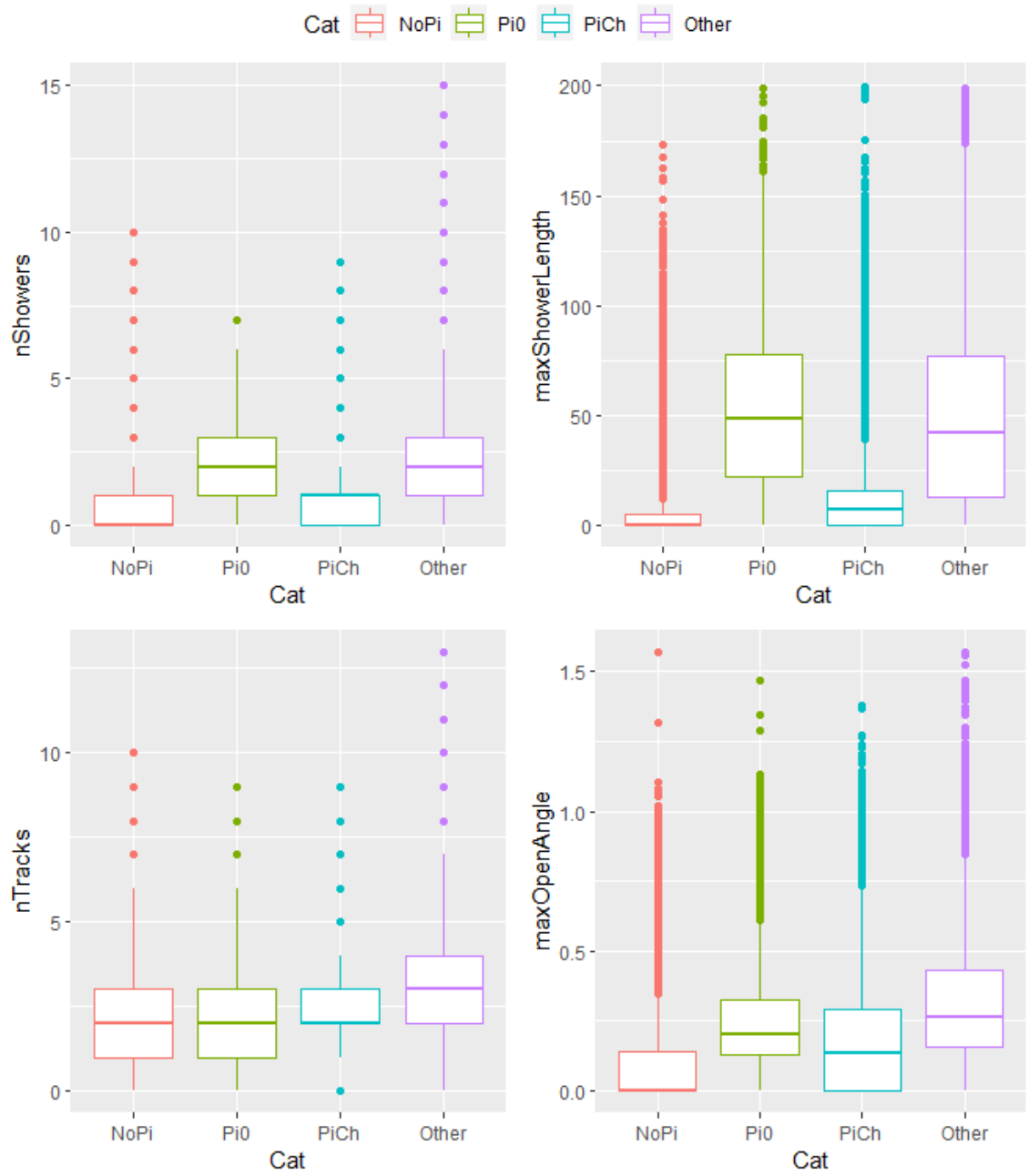

Figura 9: Diagramas de cajas por categoría para algunas variables continuas y discretas.

variables más de cerca.

4.3.1 Separación de dos variables para 2 clases

En la Figura 11, se presenta el gráfico de dispersión ampliado junto con los histogramas por categoría. También se incluyen unas elipses que ayudan a interpretar la concentración de los datos en el plano. Estas se construyen asumiendo que los datos proceden de normales bivariantes, estando centradas en sus medias y con sus ejes en las direcciones de los vectores propios de las matrices de covarianza. La longitud de los ejes es

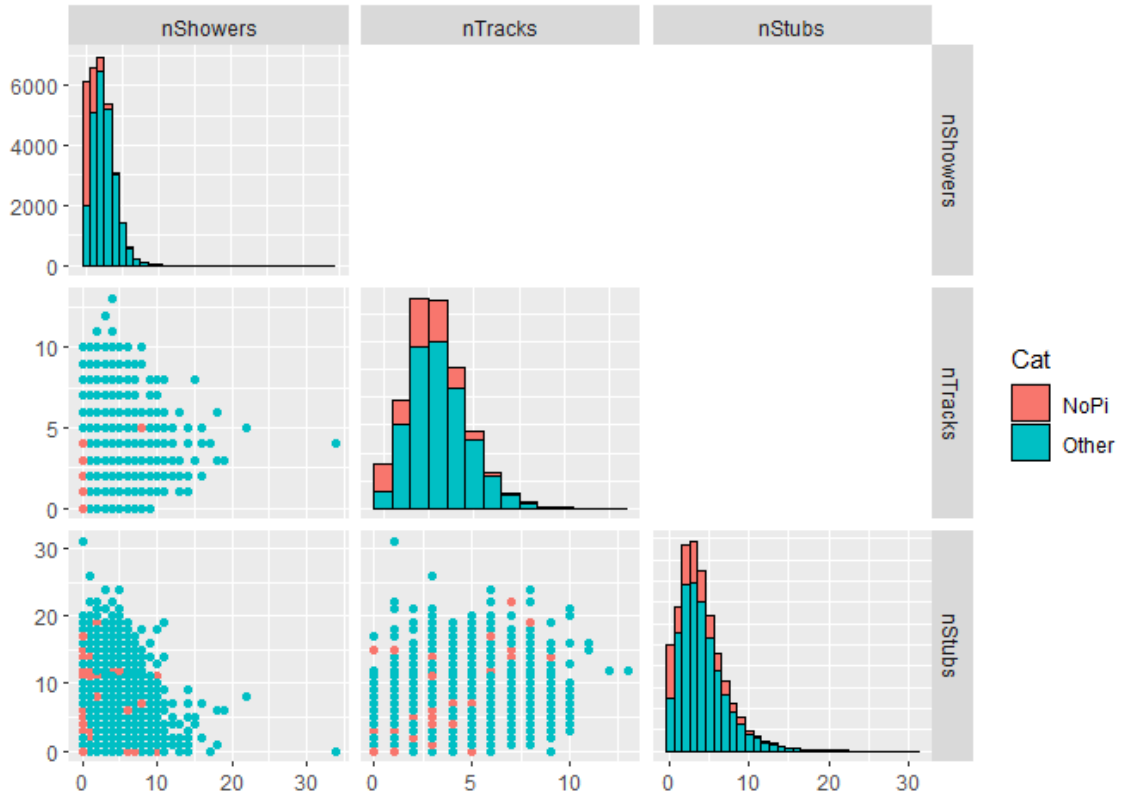

Figura 10: Gráfico de dispersión de las variables discretas.

proporcional al valor de la $\chi^2_{df=2,\alpha=0.05}$[19].

Dado que las variables son discretas, los puntos del plano serán referidos como celdas (x, y), con x las cascadas e y las trazas. Aclarado esto, se plantea como punto de partida para la clasificación un algoritmo rudimentario que, para cada celda, asigna todas las observaciones a la clase con mayor frecuencia relativa en dicha celda,

$$\max \left\{ n_j^{(x,y)} / N_j : j = NoPi, Other \right\},$$

donde $n_j^{(x,y)}$ es el número de instancias de la clase j en la celda (x, y) y N_j es el número total de instancias de la clase j. Es decir, se realiza un barrido por todas las celdas y se elige cada una del color de la clase ganadora según este criterio. Hay que notar que en la Figura 11 los colores que se muestran atienden solo a la frecuencia absoluta por celda. Así, dado que la categoría con $\geq 2\pi$ cuadruplica a la de 0π, antes de aplicar el algoritmo ya se sabe el destino de las celdas en rojo y son las celdas azules las que podrían cambiar de color si su frecuencia relativa es menor.

La precisión de este algoritmo resulta ser del 80.8 % y la matriz de confusión que lo demuestra es la de la Figura 12.

Finalmente, se presentan los valores de la eficiencia (TPR) y la pureza (PPV), sus contrapartes para negativos, especificidad (TNR) e índice predictivo negativo (NPV), y los coeficientes F_1 y κ:

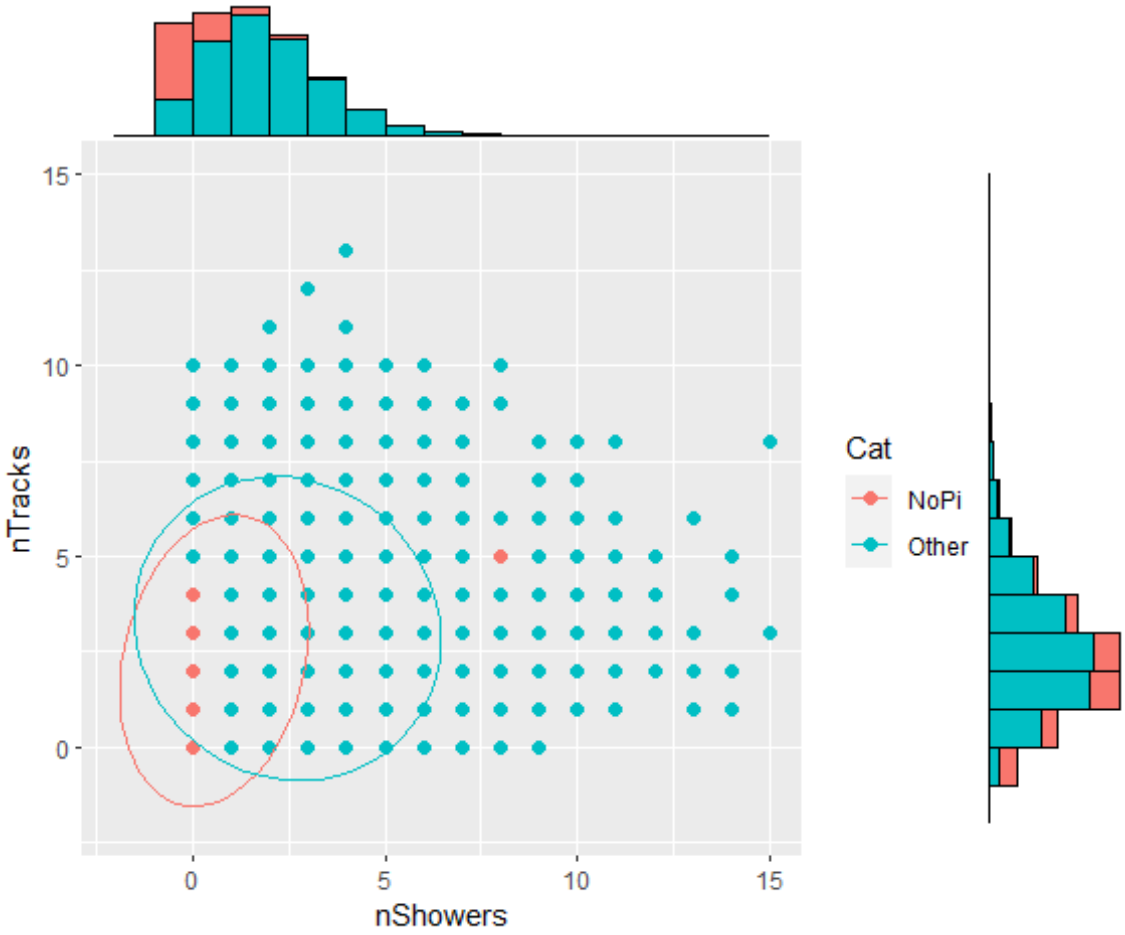

Figura 11: Gráfico de dispersión de cascadas frente a trazas.

$TPR = 81.6\,\%$	$PPV = 52.5\,\%$
$TNR = 80.6\,\%$	$NPV = 94.4\,\%$
$F_1 = 63.9\,\%$	$\kappa = 51.7\,\%$

Vista la simplicidad del algoritmo, podía pensarse que el buen valor para la precisión tendría que ver únicamente con la alta prevalencia de la clase *Other*. Esto es, el algoritmo está clasificando todas las observaciones como *Other* y, puesto que son la mayoría, está obteniendo una buena precisión. Sin embargo, esto queda desmentido al comprobar que la sensibilidad y especificidad rondan también el 80 %, es decir, ambas categorías se clasifican igual de bien. El κ es también bastante satisfactorio.

4.3.2 Separación de dos variables para 4 clases

Con todas las topologías incluidas se aplica el mismo algoritmo para obtener una precisión del 49.7 %. Su matriz de confusión es la de la Figura 13.

Como antes, se presentan las métricas derivadas de la matriz de confusión:

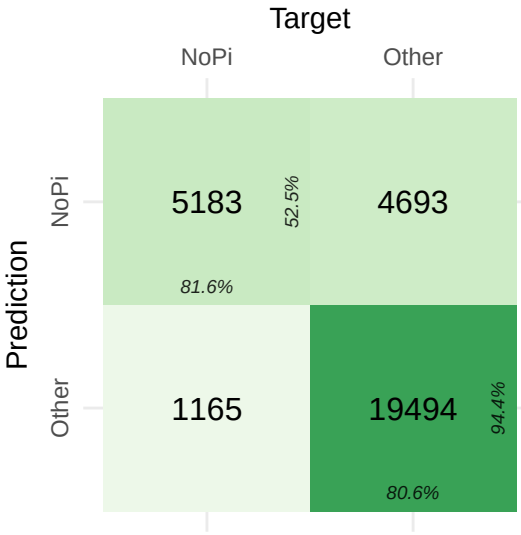

Figura 12: Matriz de confusión de 2 topologías para el algoritmo de barrido. En la diagonal se muestran los porcentajes que representa la casilla en su fila (a la derecha) y en su columna (abajo).

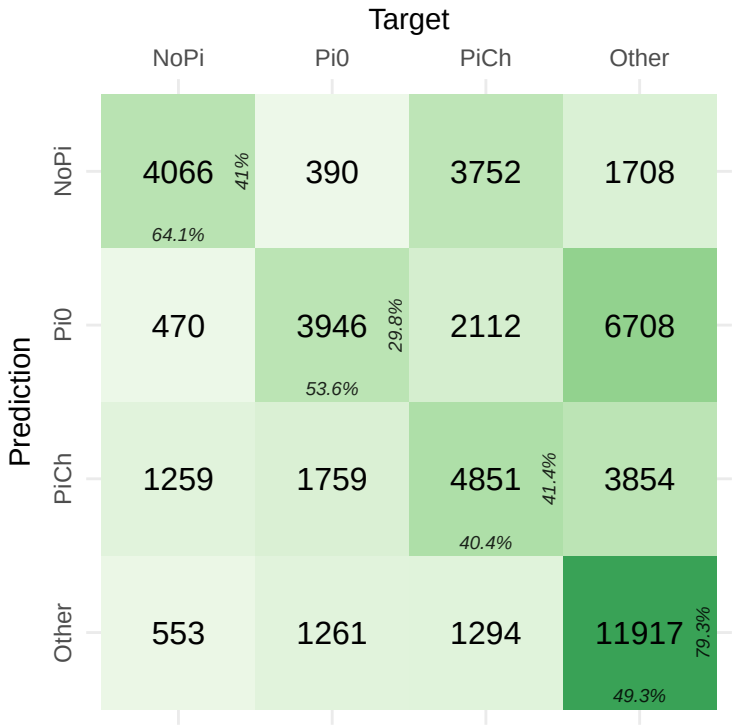

Figura 13: Matriz de confusión de 4 topologías para el algoritmo de barrido.

$TPR = 51.8\,\%$	$PPV = 47.9\,\%$
$TNR = 83.6\,\%$	$NPV = 82.8\,\%$
$F_1 = 47.5\,\%$	$\kappa = 31.2\,\%$

Claramente, al aumentar en número de topologías la complejidad del problema aumenta sobremanera y lo que antes daba buenos resultados ahora está lejos de lo deseable. Estos valores de las métricas constituyen el punto de partida del análisis y son el objetivo a superar cuando se apliquen técnicas multivariantes.

4.4 Resultados del Aprendizaje Supervisado

4.4.1 Análisis Discriminante

Tal y como se ha indicado a lo largo del trabajo, el análisis discriminante, tanto lineal como cuadrático, es un método de clasificación adecuado para el objetivo que perseguimos en esta aplicación práctica. Si bien, requiere de unos supuestos para su aplicación que habría que tener en cuenta. En primer lugar necesita satisfacer que el vector formado por todas las variables explicativas tenga una distribución de probabilidad normal multivarinate. Para este experimento, este supuesto no se puede asumir puesto que algunas de las variables explicativas no son ni tan siquiera variables continuas. Además en el caso del análisis discriminante lineal se necesita verificar el supuesto de homogeneidad (igualdad) de varianzas en cada uno de los niveles de la variable respuesta. Si bien, es ampliamente aceptado en el ámbito de la Ciencia de Datos y el Machine Learning que para conjuntos de datos con un número muy elevado de registros, existe cierta robustez de estos clasificadores ante la ausencia de estos supuestos. Es por ello que se han estimado los coeficientes de los dos modelos de clasificación discriminante lineal y cuadrático considerando como variable respuesta la diseñada para este experimento con cuatro niveles, y como variables explicativas las variables C, T, S, LC, LT y AC. Se ha hecho una validación cruzada obteniendo una tasa de clasificaciones correctas tanto en el caso del discriminante lineal como en el del discriminante cuadrático no excesivamente elevada. Las figuras 14 y 15 siguientes muestran la validación realizada con cada uno de los métodos.

La utilidad de estos modelos no parece mucha en este caso. Es muy probable que la baja tasa de clasificaciones correctas sea debida a que la variable respuesta no es binaria. Es por esto que una práctica habitual justificada es ajustar modelos binarios que comparen todos los niveles de la variable respuesta con un nivel de referencia de la misma.

Se ilustrará este enfoque para el caso de la regresión logística en la sección siguiente.

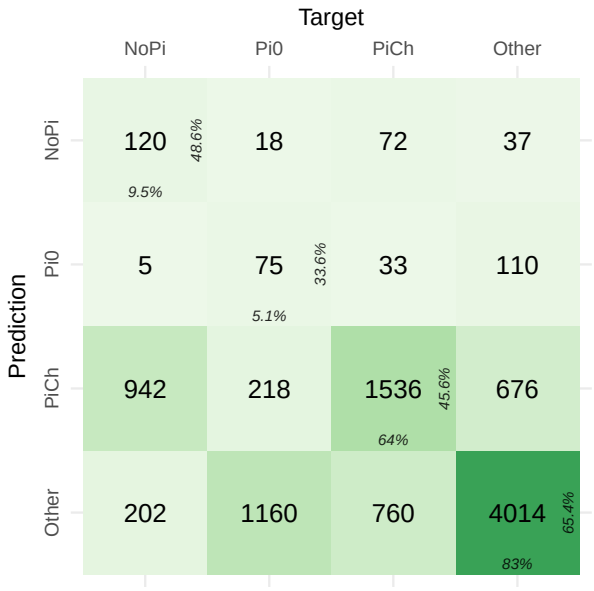

Figura 14: Matriz de confusión modelo LDA.

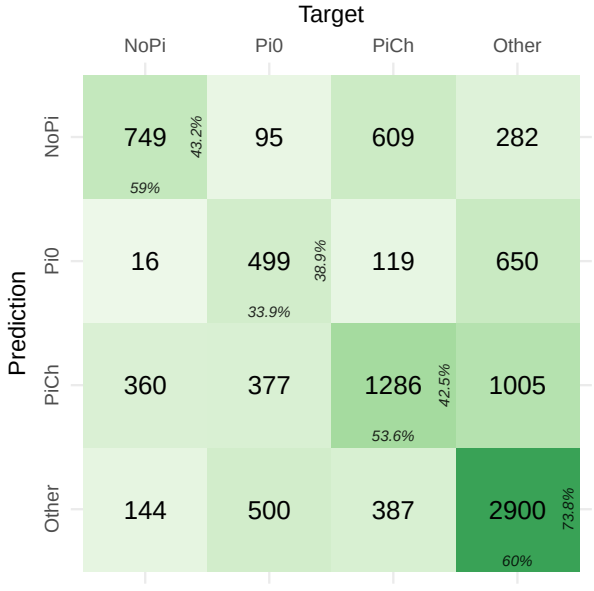

Figura 15: Matriz de confusión modelo QDA.

4.4.2 Regresión Logística

En esta sección se presenta la aplicación del modelo de regresión logística. Recuérdese que si son cuatro categorías entre los sucesos de piones son tres, entonces, los logits a aplicar. Cada uno devolverá las probabilidades de pertenecer a la categoría base, sin piones, frente a una de las demás, que todas tienen piones. El proceso de selección del modelo se realiza mediante el método *stepwise forward*, que indica cuáles de las variables explicativas son estadísticamente relevantes en cada caso.

NoPi vs Pi0 El modelo estimado de regresión logística binaria para discriminar entre topologías sin piones en el estado final y aquellas con un pion neutro tiene la siguiente forma:

$$\widehat{L}_{Pi0} = \widehat{\alpha} + \widehat{\beta}_C C + \widehat{\beta}_T T + \widehat{\beta}_S S + \widehat{\beta}_{LC} LC. \tag{4.1}$$

Los parámetros del modelo se encuentran en la Tabla 3.

Tabla 3: Coeficientes del modelo *NoPi* vs *Pi0*.

Variable	β	**DT**	**Z**	**p**	**OR**
Constante	-1.285	0.056	-23.083	< 0.001	
C	0.847	0.033	25.384	< 0.001	2.332
T	-0.110	0.024	-4.515	< 0.001	0.896
S	-0.146	0.013	-11.601	< 0.001	0.864
LC	0.046	0.001	31.605	< 0.001	1.047

β = parámetros estimados, DT = desviación típica,
Z = estadístico Z, p = p-valor, OR = odds ratio.

A la vista de los resultados obtenidos para el test de Wald, que contrasta si cada uno de los parámetros asociado de forma individual a cada variable explicativa que ha entrado en el modelo, es o no significativamente distinto de cero a nivel poblacional, se puede afirmar que todas lo son puesto que p-value < 0.001 en todos los casos. Esto permite un uso más allá de tener un mejor o peor modelo de pronóstico para la clasificación en un nivel u otro de la variable respuesta. En efecto, los modelos de regresión logística tienen la particularidad de que, en el caso de ser significativas, nos permiten modular la fuerza con la que queda variable explicativa influye en la clasificación. Esta información se tiene de la odd-ratio o cociente de ventajas. La exponencial de un parámetro del modelo es el cociente de ventajas de respuesta *Pi0* frente a la *NoPi* para incrementos unitarios en la variable explicativa que se esté interpretando. Así, un incremento unitario en el número de cascadas hace 2.33 veces más posible que el evento haya tenido un π^0 en lugar de no haber tenido ninguno. Mirando ahora a la OR de la variable LC, se aprecia que el efecto de aumentar una unidad es mínimo, pero la interpretación se extiende a cualquier cambio en la variable. Por ejemplo, si la longitud de la cascada más larga aumentara en 15 cm, entonces es casi el doble de posible, $\left(e^{0.046}\right)^{15} = 1.99$, que haya habido un π^0 en el estado final frente a que no haya habido ninguno.

Para finalizar, la Figura 16 es la matriz de confusión incluyendo algunas métricas de diagnóstico. Es grato obtener una buena sensibilidad para los sucesos sin piones. Recuérdese que se trata del canal más limpio y mejor conocido teóricamente, con lo que es importante aislarlo. El porcentaje de clasificaciones correctas y otras métricas son:

$ACC = 86.6\,\%$	$F_1 = 86.0\,\%$	$\kappa = 73.2\,\%$

NoPi vs PiCh Si se aplica ahora la regresión logística a la categoría base frente a la de un $\pi^{\pm}$ todas las variables son incluidas:

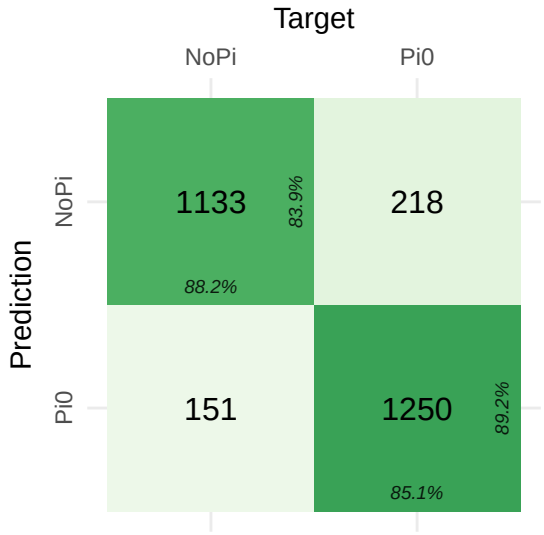

Figura 16: Matriz de confusión *Nopi* vs *Pi0* para regresión logística. Los porcentajes en la diagonal superior son *Sensibilidad* (abajo) y *PPV* (a la derecha) y en la diagonal inferior *Especificidad* (abajo) y *NPV* (a la derecha).

$$\widehat{L}_{PiCh} = \widehat{\alpha} + \widehat{\beta}_C C + \widehat{\beta}_T T + \widehat{\beta}_S S + \widehat{\beta}_{LC} LC + \widehat{\beta}_{LT} LT + \widehat{\beta}_{AC} AC. \tag{4.2}$$

Las estimaciones del modelo se pueden consultar en la Tabla 4. Tal y como sucedía en el modelo logit anterio, a la vista de los resultados del test Wald, todas las variables que han entrado en el modelo son significativas a nivel poblacional (p-value < 0.001) de modo que también tiene sentido la interpretación individual de sus cocientes de ventaja. Aunque, en esta ocasión parece que no hay mucha interpretabilidad física en términos del cociente de ventajas. Podría pensarse que los incrementos en AC van a marcar grandes diferencias ($OR = 4.12$) pero en realidad esta variable no va a cambiar más que un par de décimas, en cuyo caso la posiblidad de tener un $\pi^{\pm}$ frente a no tener ninguno aumenta solamente $4.12^{0.2} = 1.33$ veces. Si se mira a S, tiene un valor negativo al igual que en el modelo anterior y por tanto su odds ratio es menor que uno. Entonces, por cada incremento unitario en S, es $1/0.866 = 1.16$ veces más posible que el suceso no haya tenido piones frente a que haya tenido un $\pi^{\pm}$. Similarmente ocurre con un π^0. ¿Son acaso los *stubs* una manera de medir la resistencia a la aparición de piones? A menudo un *stub* es una cascada o traza mal reconstruida, que no se identifica como tal. Que haya pues sucesos *NoPi* con más *stubs* indica que probablemente en realidad había un pion que no ha dejado una señal suficientemente clara en el detector.

La Figura 17 es la matriz de confusión y, como antes, se presentan también las siguientes métricas:

$ACC = 72.9\,\%$	$F_1 = 49.1\,\%$	$\kappa = 32.9\,\%$

NoPi vs Other Se completa la regresión logística con el modelo entre las dos clases más extremas. Esta vez solo queda fuera la longitud de la traza más larga, expresándose

Tabla 4: Coeficientes del modelo *NoPi* vs *PiCh*.

Variable	β	**DT**	**Z**	**p**	**OR**
Constante	-0.1423	0.0412	-3.454	< 0.001	
C	0.3629	0.0354	10.254	< 0.001	1.4375
T	0.3121	0.0166	18.753	< 0.001	1.3662
S	-0.2004	0.0084	-23.909	< 0.001	0.8184
LC	0.0077	0.0012	6.174	< 0.001	1.0077
LT	0.0008	0.0002	4.298	< 0.001	1.0008
AC	1.4161	0.1383	10.242	< 0.001	4.1210

β = parámetros estimados, DT = desviación típica,
Z = estadístico Z, p = p-valor, OR = odds ratio.

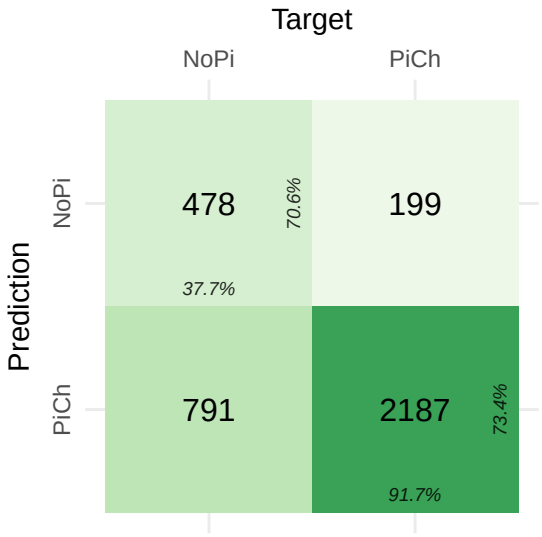

Figura 17: Matriz de confusión *Nopi* vs *PiCh* para regresión logística.

el logit:

$$\widehat{L}_{Other} = \widehat{\alpha} + \widehat{\beta}_C C + \widehat{\beta}_T T + \widehat{\beta}_S S + \widehat{\beta}_{LC} LC + \widehat{\beta}_{AC} AC. \tag{4.3}$$

Los parámetros estimados están en la Tabla 5 y las medidas de diagnóstico en la Figura 18. Además, se tiene:

$$ACC = 88.0\,\% \quad F_1 = 69.2\,\% \quad \kappa = 61.8\,\%$$

Para este modelo también se tiene la particularidad de que cada variable explicativa que ha entrado en el modelo es significativa (p-value < 0.001) y, por tanto, tienen sentido interpretar sus odds. Como cabía esperar, aumentos en el número de cascadas y trazas favorecen a la categoría *Other*. Se calcula el efecto combinado sobre las odds de aumentar C y T en una unidad como $e^{0.810}e^{0.550} = 3.9$, es decir, se cuadruplica la posibilidad tener más de un pion en el estado final frente a no tener ninguno.

Tabla 5: Coeficientes del modelo *NoPi* vs *Other*.

Variable	β	DT	Z	p	OR
Constante	-1.666	0.049	-34.031	< 0.001	
C	0.810	0.031	26.419	< 0.001	2.247
T	0.550	0.016	34.776	< 0.001	1.734
S	-0.143	0.008	-17.239	< 0.001	0.866
LC	0.033	0.001	26.933	< 0.001	1.034
AC	1.017	0.132	7.676	< 0.001	2.764

β = parámetros estimados, DT = desviación típica,
Z = estadístico Z, p = p-valor, OR = odds ratio.

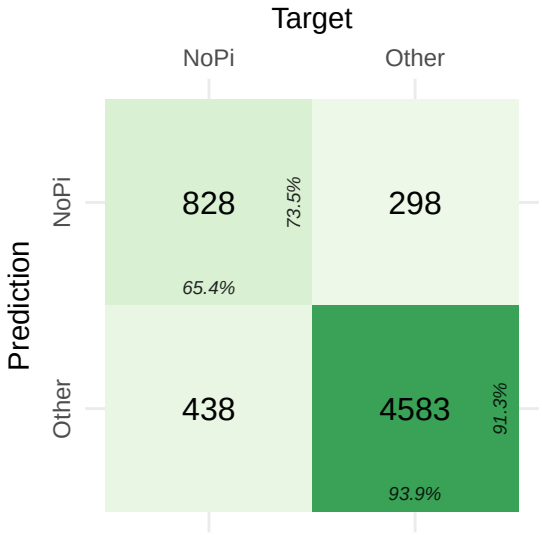

Figura 18: Matriz de confusión *Nopi* vs *Other* para regresión logística.

4.4.3 K-Nearest Neighbours

Finalizando con los métodos de clasificación, se presenta la aplicación del algoritmo de los k vecinos más cercanos. Antes de los resultados, dos consideraciones importantes. Primero, recordar la importancia de tipificar el dataset. Se han hecho pruebas con los datos en crudo y se pierde en torno al 5 % de eficiencia. Después, unas palabras sobre el (hiper)parámetro k. El desempeño del algoritmo es totalmente dependiente de su elección: valores muy pequeños de k proporcionan un ajuste muy flexible a costa de alta variabilidad y riesgo de *overfitting* mientras que valores grandes pueden enmascarar la estructura del conjunto de datos. Se trata de una clara manifestación del *bias-variance tradeoff*[g]. En la práctica hay muchas maneras de ajustar k, siendo la aquí seguida la validación cruzada con 10 hojas (*10-fold CV*): se divide el conjunto de entrenamiento en diez subconjuntos y, secuencialmente, se ajusta el modelo con nueve de ellos y se estima

[g]El *bias* o sesgo se refiere a las suposiciones simplificadas que hace el modelo y la varianza se refiere a la sensibilidad del modelo a las fluctuaciones en los datos. El *trade-off* entre sesgo y varianza implica encontrar un equilibrio adecuado para obtener un modelo que sea lo suficientemente flexible para capturar la complejidad de los datos sin sobreajustarse a ellos.

el error con el restante. Es el método más costoso computacionalmente de los utilizados pero proporciona una estimación confiable del rendimiento del modelo en el conjunto de test.

Comentar que, en este caso, el enfoque de pesar los sucesos según la expresión 3.38 no ha tenido éxito. Esencialmente, devuelve una clasificación aleatoria. Por tanto, se ha ejecutado el método sin ninguna compensación para las categorías. Se muestran en la Figura 19 los resultados para $k = 75$, elegido mediante CV con el criterio de maximizar el coeficiente κ. La tasa de clasificaciones correctas es la mejor que se ha podido obtener al considerar los cuatro niveles de respuesta:

$ACC = 62.1\,\%$	$F_1 = 52.9\,\%$	$\kappa = 38.7\,\%$

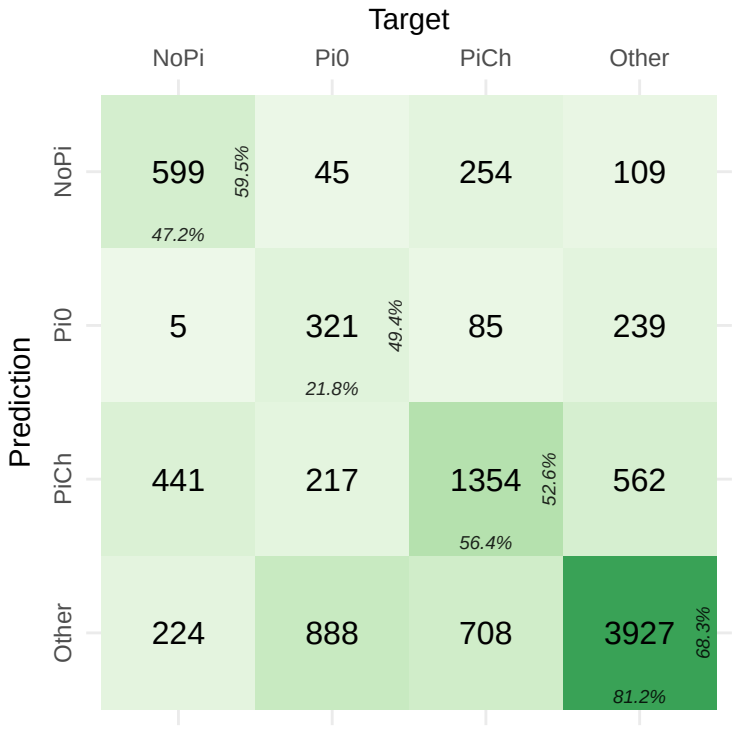

Figura 19: Matriz de confusión para el algoritmo kNN con $k = 75$.

5 Conclusiones

En la línea de lo expuesto en la Introducción, se resume el trabajo que se ha llevado a cabo y se evalúa si se han cumplido los objetivos. Desde el punto teórico, los contenidos se han logrado desarrollar de manera breve pero necesaria para ganar un entendimiento suficiente del experimento, sus objetivos y sus implicaciones y poder llevar a cabo la clasificación. Personalmente, me quedo con la satisfacción de haber profundizado por primera en el funcionamiento de un proyecto científico de la envergadura de SBN y haber descubierto la complejidad y el esfuerzo colaborativo que representa. También, haber entendido el principio de las LAr-TPCs y su importancia en el futuro de la física de neutrinos. A destacar en la parte matemáticas que, a pesar de haberse fundamentado algunos de los modelos más sencillos, estos son una gran base introductoria y fundamental para seguir aprendiendo en el campo del Aprendizaje Automático.

Pasando a la aplicación, es necesario resaltar en primer lugar la dificultad del problema considerado. Las topologías finales que se han querido clasificar sufren gran solapamiento, inherente a cómo han sido definidas. Además, el hecho de tener categorías descompensadas en número de sucesos supone un inconveniente adicional. Sin embargo, ello es propio de la naturaleza del experimento y se ha tenido que acatar así. En cuento a los modelos, ha quedado claro que las técnicas de análisis discriminante empleadas no son apropiadas para una clasificación de este tipo y no han aportado mucho valor a la discusión. Por el contrario, si bien la regresión logística tampoco ha devuelto unas clasificaciones excepcionales en cuanto a métricas, es de valorar muy positivamente la riqueza interpretativa que se ha extraído de las odds ratio. Por último, el algoritmo de los k vecinos más cercanos ha servido para iniciarse en métodos más computacionales y en las sutilezas que les envuelven, como la optimización de los hiperparámetros del modelo o el riesgo de overfitting.

A fin de ser crítico, hay que reconocer que para abordar un problema de esta complejidad de manera más correcta se necesita más planificación en varios sentidos. Por una parte, en cuanto al trabajo previo a lanzar algoritmos, una depuración más profunda de los datos así como una correcta gestión de los muy presentes valores extremos es necesaria. Esos han perjudicado la capacidad de los modelos para adaptarse a los datos. En el otro extremo ha quedado claro que hay mucho por hacer también. Es decir, se han aplicado técnicas multivariantes cuyo contenido matemático es abarcable en un Trabajo de Fin de Grado, y han sido ricas en interpretación y en ocasiones incluso han dado clasificaciones medianamente buenas. No obstante, existen muchas herramientas más complejas que pueden dar resultados mejores, y no han podido ser abordadas aquí.

Queda como trabajo futuro, pues, la tarea de explorar otras posibilidades dentro del Machine Learning para su aplicación a este proyecto y a otros muchos de la Física de Partículas. Siendo que la cantidad de datos recogida en los experimentos no para de crecer y que estos son el combustible para los algoritmos de aprendizaje, estoy seguro de que estas dos disciplinas han quedado ya unidas para la eternidad.

Referencias

[1] M. Thomson, *Modern particle physics*. New York: Cambridge University Press, 2013.

[2] P. A. Machado, O. Palamara, and D. W. Schmitz, "The short-baseline neutrino program at fermilab," *Annual Review of Nuclear and Particle Science*, vol. 69, pp. 363–387, oct 2019.

[3] D. Zelterman, *Applied Multivariate Statistics with R*. Springer Cham, 2016.

[4] A. Agresti, *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics, Wiley, 2015.

[5] A. M. Aguilera del Pino, *MODELOS DE RESPUESTA DISCRETA*.

[6] E. Vera-Salmerón, E. Mota-Romero, J. L. Romero-Béjar, C. Dominguez-Nogueira, and B. Gómez-Pozo, "Pressure ulcers risk assessment according to nursing criteria," *Healthcare*, vol. 10, no. 8, 2022.

[7] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, fourth ed., 2002. ISBN 0-387-95457-0.

[8] Kuhn and Max, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, p. 1–26, 2008.

[9] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[10] S. Fukuda and Y. Fukuda, "The super-kamiokande detector," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 501, no. 2, pp. 418–462, 2003.

[11] A. Bellerive, J. Klein, A. McDonald, A. Noble, and A. Poon, "The sudbury neutrino observatory," *Nuclear Physics B*, vol. 908, pp. 30–51, jul 2016.

[12] "Sitio web particle data group." https://pdg.lbl.gov/.

[13] A. A. Aguilar-Arevalo, "Event excess in the MiniBooNE search," *Physical Review Letters*, vol. 105, oct 2010.

[14] F. N. A. Laboratory, "Sitio web sbnd." https://sbn-nd.fnal.gov/.

[15] C. Rubbia, "The liquid-argon time projection chamber: A new concept for neutrino detectors," *CERN EP Internal Reports*.

[16] A. J. Izenman, *Modern Multivariate Statistical Techniques*. New York: Springer New York, 2008.

[17] D. J. Olive, *Robust Multivariate Analysis*. Springer Cham, 2017.

[18] R. F. Casal, "Aprendizaje estadistico." https://rubenfcasal.github.io/aprendizaje_estadistico/.

[19] P. S. D. of Statistics, "Applied multivariate statistical analysis." https://online.stat.psu.edu/stat505.